RICE UNIVERSITY

Inferring Spectral and Spatiotemporal Dependencies from Data and its Application to Epilepsy

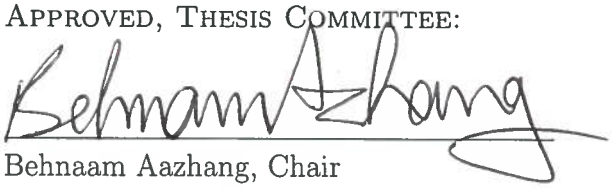
by

Rakesh Malladi

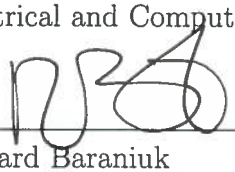
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

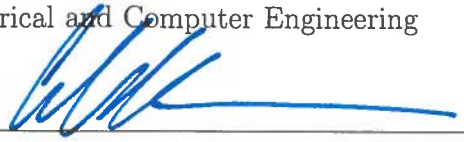
Approved, Thesis Committee:



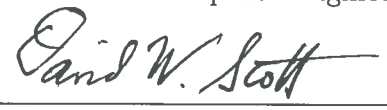
Behnaam Aazhang, Chair
J.S. Abercrombie Professor
Electrical and Computer Engineering



Richard Baraniuk
Victor E. Cameron Professor
Electrical and Computer Engineering



Caleb Kemere
Assistant Professor
Electrical and Computer Engineering



David W. Scott
Noah Harding Professor
Statistics

Houston, Texas
March, 2017

ABSTRACT

Inferring Spectral and Spatiotemporal Dependencies from Data and its Application to Epilepsy

by

Rakesh Malladi

A fundamental problem in many science and engineering disciplines is inferring the characteristics of a physical or biological system from the dependencies in data recorded from the system. The dependencies in data, particularly in case of signals recorded from brain, are commonly believed to be nonlinear and the underlying model is often unknown. This thesis focusses on developing novel information-theoretic approaches to detect and quantify spectral and spatiotemporal dependencies from data in a data-driven manner and applies them to electrocorticographic (ECoG) recordings from epilepsy patients to unravel epileptic brain networks. Frequency components in a signal or between two signals, not necessarily at the same frequency, are spectrally dependent if they are not statistically independent. Two signals are temporally dependent if the past measurements at one decrease the uncertainty in predicting the other.

First, we define a novel metric, mutual information in frequency, to detect spectral dependency and quantify it using a data-driven estimator. We then develop a data-driven estimator of mutual information between dependent data using mutual information in frequency. Next, we develop a model-based and a data-driven estimator of directed information to detect and quantify the temporal dependencies in data.

Finally, we apply the proposed metrics to ECoG recordings from epilepsy patients to identify seizure onset zone (SOZ), to learn the spatiotemporal characteristics of seizures and to infer the cross-frequency coupling in SOZ. We observe that seizure onset zone drives the rest of brain to a seizure during pre-seizure and seizure periods, while it acts as a sink during post-seizure periods. In addition, high frequency coupling increases during seizures within an ECoG channel and between channels in the same anatomical region in SOZ, but not between different regions in SOZ. This suggests different anatomical regions in the SOZ are independently driving the seizure activity and any treatment should potentially target these regions simultaneously. Going forward, the dependencies unraveled by the proposed metrics should be further analyzed to optimize the parameters of closed-loop electrical stimulation based treatments for epilepsy.

Acknowledgments

I would like to express my thanks to all the people who have helped me and made this thesis possible. First and foremost, I would like to express my sincere gratitude to my advisor Prof. Behnaam Aazhang for being an excelled advisor. He gave me time, encouraged me to discover and chart my own course, after I switched to neuroengineering in my second year. Thank you for giving me this opportunity and believing in me. I would also like thank Dr Giridhar Kalamangalam and Dr Nitin Tandon for introducing me to the world of epilepsy and for providing me with the clinical data sets, without which this thesis would not be possible. I would also like to thank Prof Don Johnson for helping me when I showed up in his office with difficult theoretical questions and guiding me towards solving them. I would like to thank Prof Caleb Kemere, Prof Jacob Robinson and Prof Xaq Pitkow for the wonderful intellectual discussions I had with them about neuroengineering. I would like to thank Prof Richard Baraniuk, Prof Caleb Kemere and Prof David Scott for being on my committee and for their time and support. Their questions, criticism and suggestions played an important role in the development of this thesis. Finally, I would like to thank Prof Ashutosh Sabharwal for imparting his wisdom, particularly while we were working late at night in Duncan.

To all my fellow Ph.D. students at Rice, thank you for filling my time here at Rice with wonderful memories. I would like to specially acknowledge Samantha, Achal, Brett, Corina, David, for teaching me how to deliver a great presentation. I would like to thank Nancy, Mayank, Rajoshi, Achal and Samantha for the numerous technical discussions we had. I would like to thank Sugnaya for providing a neuroscientist's perspective of my work. I would like to thank all other my group members, Zhiting,

Sudha, Joe, Shi, Boqiang, Negar and Paz, for making me feel at home.

I also had the great group of friends at Rice, Mehul, Kaushik, Satyakam, Kuldeep, Vaideesh, Suguman, Himanshu, Yaswanth, Adithya, who made my life in Houston enjoyable. In addition, I would like to thank my undergrad friends, especially, Praneeth and Sujay, for the numerous phone conversations and their encouragement over the last six years.

Finally, I would like to thank my mother and father, for letting me pursue my own dreams. Thank you for your love and unconditional support.

Contents

Illustrations

Tables

Chapter 1

Introduction

1.1 Learning Dependencies from Data

Inferring the dependencies from data recorded using a wide variety of sensors is a fundamental problem in many science and engineering disciplines. The objective is to learn the underlying statistical model of the system from the recorded data. If the data samples recorded from these sensors are independent and identically distributed (i.i.d.) over time, then each sensor can be modeled using a random variables and we can learn the joint distribution over all the random variables from i.i.d. data either using a model-based approach or a data-driven approach. Model-based approaches assume that the data is sampled from a parametric family of distributions and learn the parameters of the model from data, typically using one of Bayesian [1] and optimization [2] based approaches. Some of the classical models used are Gaussian graphical models, Markov models and Bayesian networks [3, 4]. Data driven approaches do not make any parametric model assumptions and only impose some smoothing assumptions. The joint distribution is typically learnt using histograms and kernel density estimation [5]. However, data samples are not i.i.d. in many applications and the aforementioned methods are not readily applicable. Developing algorithms to learn relationship structure from non-i.i.d. data is the focus of this thesis.

The data samples could be non-i.i.d. either because the underlying model is time-varying, but independent across time or identically distributed, but dependent

across time or both. Except for a few studies focussed on time-varying Gaussian model [6], Markov random fields [7] and few others, not much work was done on learning time-varying distributions. In this thesis, we do not focus on data sampled from time-varying distributions. We assume that the changes in the underlying state of the system are algorithmically identified using change point detection algorithms [8] or dealt with using sliding windows. We developed an online Bayesian change point detection algorithm to learn the change points [9] and it is described in Appendix A. This thesis focusses on learning the dependence relationships from dependent data, assuming the data is stationary in a small window and tackles the changes in the underlying system state using sliding windows.

Learning the joint distribution of dependent data is a non-trivial problem in general. This problem can be solved if the data is modeled by specific families of parametric distributions like multivariate autoregressive (MVAR) model [10], dynamic causal models [11], dynamic Bayesian networks [12]. However, in most real-world applications, the underlying model is unknown and it is often is non-linear and non-Gaussian. This is especially true in case of data recorded from brain, since we don't need billions of neurons in the brain if the brain is a linear processing machine. We therefore, focus on learning the structure from dependent data using data-driven approaches, without making any parametric assumptions. Learning the joint distribution across all sensors from dependent data in a data-driven manner is not feasible due to the curse of dimensionality [13, 14]. Instead, we focus on learning pairwise structure between data in a data-driven manner and consider three different, but very closely related characterizations of the dependence relationships between data:

1. Detecting if two data streams are statistically independent or not, and if not quantify the dependence.

2. Detecting and quantifying the statistical dependence between various frequency components within and between data streams.

3. Detecting and quantifying the causal connectivity between data streams using directed information.

We developed a novel information-theoretic metric, mutual information in frequency (MI-in-frequency), to quantify the statistical dependence in frequency [15–17] and used MI-in-frequency to estimate mutual information (MI) between the dependent data non-parametrically [16, 17]. We developed model-based and data-driven estimators of directed information to infer the causal connectivity between the data streams [18–20]. We then use the novel information-theoretic metrics developed to unravel epileptic brain networks and the motivation to apply these metrics to recordings from epilepsy patients is described in the following section.

1.2 Closed Loop Neuromodulation for Epilepsy

Epilepsy is a very common neurological disease characterized by repeated and unprovoked seizures–periods in which hyper-synchronous neural activity spreads from one or more small diseased circuits in the brain to malignantly entrain activity more broadly. Epilepsy affects nearly 1% of the world's population and there are about three million epilepsy patients in United States alone. Nearly one-third of all epilepsy patients suffer from medically refractory epilepsy (medication using anticonvulsant drugs is not effective in these patients). A substantial portion of the medically refractory epilepsy patients have focal epilepsies–their seizure networks are spatially localized to a small cortical region. In these cases, surgical resection of the epileptogenic focus can potentially cure their seizures. The focal zone that needs to be resected

is identified from the electrocorticographic (ECoG) signals recorded from the arrays of electrodes implanted in these patients. A majority of patients with implanted electrodes and an identifiable focus can be resected with only a decent probability of improved seizure outcome [21, 22]. However, permanent surgical resection risks damage to critical functional zones that are frequently adjacent or even overlapping with the seizure focus. This is especially true for epilepsy in which seizures originate in the mesial temporal lobe of the language dominant hemisphere. In these cases, resection of the seizure zone may lead to a significant decline in verbal memory [23]. In addition, a fraction of patients have multifocal epilepsies that are not well suited to resective surgeries. In these cases, and more broadly for all patients with epilepsy, an ideal solution might be a neuromodulation strategy in which stimulation is used to induce plasticity that serves to weaken the connectivity in the seizure network, leading to a cessation of seizures. The fact that these patients are already implanted with ECoG arrays capable of both monitoring and manipulating patterns of neural activity in complex spatiotemporal patterns, presents an opportunity for a major paradigm shift in how electrical stimulation is used to treat neurological disorders.

Buoyed by the success of electrical stimulation in treating movement disorders like Parkinson's disease, neuromodulation via electrical stimulation is considered a promising approach to treat epilepsy in patients, where the current treatment options are not effective [24]. However, unlike Parkinson's where GPi and STN are the targets of electrical stimulation [25], the epileptic disease network is unknown and patient specific. Learning the underlying epileptic circuit generating mechanisms in each patient is a crucial first step towards development of effective treatments for epilepsy [26, 27]. Some of the essential questions we need to answer to learn the epileptic network are

1. Identifying the regions of brain responsible for initiating seizures, referred to as seizure onset zone (SOZ) [28]

2. Characterizing the spread of seizure activity from SOZ to other brain regions

3. Characterizing the cross-frequency coupling in SOZ that is responsible for the hyper-synchronous seizure activity

In this thesis, we developed novel quantitative metrics and applied them to the ECoG recordings from focal epilepsy patients to answer the above three questions. Our hypothesis is that once the epileptic brain network is understood and well characterized, we can reverse engineer and find the optimal stimulation parameters to terminate epileptic seizure activity with minimal side-effects.

Related Work

Over the last few years, the problem of identifying seizure networks has been the focus of many studies in the scientific community [29–39]. Epilepsy is shown to be a dynamic disease in which the brain transitions between different states [40]. In a related body of work on brain connectivity, various models at different scales have been introduced that capture anatomical, functional, and effective connectivity of the brain (see [41–44] and references therein). A primary objective of many of these studies is to illustrate techniques that would provide insight into identifying epileptogenic zones in the resting state. Some of these studies also verify the differences in functional connectivity between normal controls and epileptic patients [35]. In addition, numerous studies have developed automated algorithms to identify the seizure onset zone from electrophysiological recordings obtained from the brains of epileptic patients [45–51]. These studies typically use metrics based on multivariate autore-

gressive (MVAR) models [50, 52] or metrics estimated in a data-driven manner [46]. Some studies also focus on the changes in connectivity and the influence of seizure onset zone before, during and after a seizure [53]. Finally, a special issue in nature neuroscience highlights the recent advances in elucidating the pathogenic events, the molecular level and the circuit level changes driving the epileptic activity (see [54] and the articles published in this issue).

1.3 Contributions of this Thesis

In this thesis, we developed novel information-theoretic approaches to learn spectral and spatiotemporal structure from data and apply them to unravel epileptic brain networks. The main contributions of my thesis are summarized below:

1. Defined a novel metric, mutual information in frequency, to measure cross-frequency coupling in data and developed a kernel density based and a nearest neighbor based data-driven algorithm to estimate MI-in-frequency.

2. Developed a data-driven algorithm to estimate mutual information between dependent data. The key idea is that MI estimation can be made tractable by focusing only on those frequencies that are statistically independent, which are identified by our MI-in-frequency metric.

3. Developed an almost surely convergent MVAR model-based and data-driven directed information (DI) estimator. Linear causal interactions between two time-series are quantified using the MVAR model-based DI estimator, whereas both linear and nonlinear causal interactions are quantified by the data-driven DI estimator.

4. Developed a MVAR model-based and a data-driven SOZ identification algorithm. We then studied the spatiotemporal seizure evolution mechanisms via causal connectivity graphs inferred from ECoG data during preictal, ictal and postictal periods. Finally, we used the MI-in-frequency metric to characterize the changes in cross-frequency coupling in SOZ during preictal, ictal and postictal periods.

The outline of the rest of the thesis is as follows. In chapter 2, the proposed MI-in-frequency metric and the data-driven estimators to estimate it are described. In addition, the proposed data-driven MI estimator is also described. The proposed model-based and data-driven DI estimators are described in chapter 3. In chapter 4, the information-theoretic metrics developed so far are applied to ECoG time-series recordings from epilepsy patients to identify the seizure onset zone and characterize the spectral and spatiotemporal dynamics of seizure activity during preictal, ictal and postictal periods. Finally, concluding remarks and some exciting future directions are outlined in chapter 5.

Chapter 2

Mutual Information in Frequency

2.1 Introduction

Learning and quantifying the dependence between multiple data streams plays an important role in many science and engineering applications. Mutual information (MI) [55, 56] is a powerful and well developed tool that has been used to measure the statistical dependence between data over time by a non-negative scalar [57]. Estimating mutual information from independent and identically distributed (i.i.d.) data is a well-studied problem [58]. However, real-world data is usually not independent across time and the underlying model is not known. Furthermore, there is tremendous interest in determining if the frequency components in the data are independent or not, and if not, in quantifying the dependence. This is especially important in areas like neuroscience where recent evidence suggests that coupling across frequencies observed in the recordings from the brain plays an important role in neuronal computation, learning and memory [59, 60]. The coupling or dependence across frequencies in the data will be referred to as cross-frequency coupling (CFC). The frequency dependence could be within a single data stream or between data streams, not necessarily at the same frequency. In this chapter, we define a novel metric to detect and quantify the statistical dependence across frequency between data and use this metric to measure mutual information between dependent data streams.

The dependence in frequency has been quantified using metrics like coherence for

multivariate autoregressive (MVAR) models [61]. The time-series data recorded from the brain, however, are not linearly related and cross-frequency coupling metrics like phase-amplitude, amplitude-amplitude, phase-phase coupling [62] and RV-coupling coefficient [63] are used instead to quantify the dependence across different frequencies. All these existing metrics cannot identify if two frequency components are statistically independent or not. In fact, a recent review article on CFC metrics suggests the use of cross-frequency 'correlation' instead of 'coupling' to describe CFC metrics [62]. In addition, even though the dependence between two frequencies is due to a combination of dependence across phase and amplitude, the existing CFC metrics are defined to treat them separately with no obvious way to combine them [60]. Furthermore, a list of confounds that affect the current phase-amplitude coupling metrics is provided in [62]. A more comprehensive metric that detects statistical independence would be invaluable in determining how neuronal oscillations are involved in computation, communication and learning in the brain.

The main contribution of this work is defining and exploiting a powerful new metric, referred to as mutual information in frequency (MI-in-frequency) to detect and quantify the statistical dependence between frequency components in data. The key idea in our work is to use Cramér's spectral representation [64, 65] to transform a time-domain stochastic process into a stochastic process in the frequency domain, the samples of which can be estimated at each frequency from the time-domain data samples [66]. We then define the MI-in-frequency as the MI between the Cramér's spectral representations of the two time series at the corresponding frequencies. It is well-known that for MVAR models with Gaussian noise, coherence is sufficient to detect the statistical dependence in frequency and we show for this class of models that the MI-in-frequency metric is related to coherence by a one-to-one function, implying

that our metric correctly captures the dependency in frequency. We then describe two data-driven algorithms – one based on kernel density estimation (KDMIF) and the other based on nearest neighbor estimation (NNMIF) – to estimate MI-in-frequency without assuming any parametric model of the data. We considered these two approaches, since they outperformed other approaches in estimating MI from i.i.d data and there is no clear winner between them [67, 68]. The performance of both these algorithms is validated and compared against modulation index, a commonly used cross-frequency coupling metric [59, 69], on data generated from multiple simulated models. Our results demonstrate the superiority of our metric, MI-in-frequency, over existing cross-frequency coupling metrics.

We then use MI-in-frequency to develop a data-driven estimator for MI between the data. Note that MI estimation is a solved problem if the data samples are i.i.d. [58] or if the underlying model is Gaussian [56, 70]. As mentioned earlier, real-world data is neither independent across time nor Gaussian and the underlying model is often unknown. Our data-driven MI estimation algorithm applies to dependent data, without making any parametric model assumptions. The key idea is to make the problem computationally tractable by focussing only on those frequencies in the two data streams that are statistically dependent, which can be identified by our MI-in-frequency metric. This estimator converges to the true value for Gaussian models and our simulation results demonstrate that it works well for nonlinear models.

The main contributions of this chapter are

- Defining a novel metric, MI-in-frequency, to measure cross-frequency coupling in data and developing data-driven algorithms to estimate MI-in-frequency.

- Developing a data-driven algorithm to estimate the mutual information between

dependent data.

2.2 Spectral Representation of Stochastic Processes

Two basic spectral representations are associated with a stochastic process - power spectral distribution and Cramér's representation. A detailed description of both these spectral representations is provided in [64, 65]. Consider a stochastic processes $X(t), t \in \mathbb{R}$. Let $S_X(\nu)$ for $\nu \in \mathbb{R}$ be the spectral distribution function of X and $s_X(\nu)$, its power spectral density, if it exists. The Cramér's representation of $X(t)$ and its key properties are stated in the following theorem (page 380 in [65]).

Theorem 2.1. *Let $X(t)$ be a second order stationary, mean-square continuous and zero mean stochastic process. Then there exists a complex-valued, finite-variance, orthogonal increment process $\widetilde{X}(\nu)$ in the frequency domain $\nu \in \mathbb{R}$, such that*

$$X(t) = \int_{-\infty}^{\infty} e^{j2\pi\nu t} d\widetilde{X}(\nu), \text{ with } \mathbb{E}\left[d\widetilde{X}(\nu)\right] = 0, \text{ and } \mathbb{E}\left[|d\widetilde{X}(\nu)|^2\right] = dS_X(\nu).$$

The process $\widetilde{X}(\nu) = \widetilde{X}_R(\nu) + j\widetilde{X}_I(\nu)$ satisfying the above theorem is the spectral process or the Cramér's representation of $X(t)$. $d\widetilde{X}(\nu)$ is the complex random variable representing the amplitude of oscillation in the interval from ν to $\nu + d\nu$ in $X(t)$. The integral in Theorem 3.1 is a Fourier-Stieltjes integral. Intuitively, Theorem 3.1 decomposes $X(t)$ into a mutually orthogonal increment process in the frequency domain. Furthermore, if the $X(t)$ is real-valued, then $\widetilde{X}(-\nu) = \widetilde{X}^{\star}(\nu), \mathbb{E}\left[d\widetilde{X}_R(\nu) d\widetilde{X}_I(\nu)\right] = 0$, and

$$\mathbb{E}\left[\left(d\widetilde{X}_R(\nu)\right)^2\right] = \mathbb{E}\left[\left(d\widetilde{X}_I(\nu)\right)^2\right] = \tfrac{1}{2} dS_X(\nu). \tag{2.1}$$

We have the following theorem (page 385 in [65]) for the special case of a real-valued Gaussian process $X(t)$.

Theorem 2.2. *Let $X(t)$ be a real-valued stationary, mean-square continuous Gaussian process with zero mean and power spectral distribution function $S_X(\nu), \nu \in \mathbb{R}$. Then the real and imaginary parts of its spectral process $\widetilde{X}_R(\nu)$ and $\widetilde{X}_I(\nu)$ are zero mean, mutually independent, identically distributed Gaussian processes satisfying* (2.1).

Example: Consider the zero mean stationary Gaussian process $X(t) = A\cos(2\pi\nu_0 t + \Theta)$, where A is Rayleigh random variable with parameter σ_A that is independent of Θ, which is uniform in $[0, 2\pi)$. The increments of the spectral process of $X(t)$ are all zero, except at $\nu = \pm\nu_0$, where the increment is $\frac{A}{2}\exp(\pm j\Theta)$ [65]. This implies that the sample path of the real part of spectral process $\widetilde{X}(\nu)$ has two jumps of same magnitude and direction at frequencies $\pm\nu_0$, while that of the imaginary part has two jumps of same magnitude, but opposite directions at $\pm\nu_0$. The magnitude of the jump at ν_0 in the real and imaginary parts is $\frac{A}{2}\cos\Theta$ and $\frac{A}{2}\sin\Theta$ respectively, both of which are Gaussian random variables with mean zero and variance $\frac{1}{2}\sigma_A^2$. This spectral process is intuitive because we know $X(t)$ has all its energy only at frequencies $\pm\nu_0$ and the variance of the increments of the spectral process $d\widetilde{X}(\nu)$ is equal to the differential power spectral distribution of $X(t)$ which is nonzero only at $\pm\nu_0$. We therefore expect all sample paths of the random process $\widetilde{X}(\nu)$ with non-zero probability to be constant, except for jumps at $\pm\nu_0$.

2.3 Mutual Information in Frequency

We first define MI between frequencies within a process and between processes in continuous time. We then extend this definition to discrete-time stochastic processes. Consider $d\widetilde{X}(\nu_i)$ and $d\widetilde{Y}(\nu_j)$, the increments of spectral processes or the Cramér's representation of $X(t)$ and $Y(t)$ at frequencies ν_i and ν_j respectively. Let $\mathrm{P}\big(d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i), d\widetilde{Y}_R(\nu_j), d\widetilde{Y}_I(\nu_j)\big)$ be the joint probability density of the four dimensional random vector of the real and imaginary parts of $d\widetilde{X}(\nu_i)$ and $d\widetilde{Y}(\nu_j)$. The corresponding two-dimensional marginal densities are denoted by $\mathrm{P}\big(d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i)\big)$ and $\mathrm{P}\big(d\widetilde{Y}_R(\nu_j), d\widetilde{Y}_I(\nu_j)\big)$. The MI between $X(t)$ at ν_i and $Y(t)$ at ν_j is defined as

$$\begin{aligned}\mathrm{MI}_{XY}(\nu_i,\nu_j) &= \mathrm{I}\big(\{d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i)\}; \{d\widetilde{Y}_R(\nu_j), d\widetilde{Y}_I(\nu_j)\}\big),\\ &= \mathbb{E}\left\{\log \frac{\mathrm{P}\big(d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i), d\widetilde{Y}_R(\nu_j), d\widetilde{Y}_I(\nu_j)\big)}{\mathrm{P}\big(d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i)\big)\mathrm{P}\big(d\widetilde{Y}_R(\nu_j), d\widetilde{Y}_I(\nu_j)\big)}\right\}, \end{aligned} \tag{2.2}$$

where $\mathrm{I}(\{\cdot,\cdot\};\{\cdot,\cdot\})$ is the standard mutual information between two pairs of two dimensional real-valued random vectors [56]. The MI between two different frequencies ν_i, ν_j in the same process $Y(t)$ is similarly defined as

$$\mathrm{MI}_{YY}\big(\nu_i,\nu_j\big) = \mathrm{I}\big(\{d\widetilde{Y}_R\big(\nu_i\big), d\widetilde{Y}_I\big(\nu_i\big)\}; \{d\widetilde{Y}_R\big(\nu_j\big), d\widetilde{Y}_I\big(\nu_j\big)\}\big). \tag{2.3}$$

The MI between the components of Y at frequencies $\nu_i = \nu_j = \nu$, $\mathrm{MI}_{YY}(\nu,\nu)$, is ∞, a consequence of the fact that $\big[d\widetilde{Y}_R(\nu), d\widetilde{Y}_I(\nu)\big]$ is a continuous-valued random vector whose conditional entropy is not lower bounded. Just for the convenience of representing our results, we set $\mathrm{MI}_{YY}(\nu,\nu)$ to be zero in the remainder of the paper. MI-in-frequency defined in (2.2), (2.3) is a non-negative number. If MI-in-

frequency between two frequencies is zero, then they are independent and if not, MI-in-frequency is a measure of the common information between the two frequency components. MI-in-frequency between two processes (2.2) is not symmetric in general, i.e., $\mathrm{MI}_{XY}(\nu_i, \nu_j) \neq \mathrm{MI}_{XY}(\nu_j, \nu_i)$. However, it is symmetric within a process, i.e., $\mathrm{MI}_{YY}(\nu_i, \nu_j) = \mathrm{MI}_{YY}(\nu_j, \nu_i)$.

Example: Continuing with our example in section 2.2, let $X(t) = A\cos(2\pi\nu_0 t + \Theta)$ and $Y(t) = X(t)^2$. Then $d\widetilde{Y}(\nu)$ is zero except at $\nu = 0$, where the spectral increment is $\frac{A^2}{2}$, and at $\nu = \pm 2\nu_0$, where the increment is $\frac{A^2}{4}\exp(\pm j2\Theta)$. As a result, the frequency component at $\pm\nu_0$ in X and at frequencies $\{0, \pm 2\nu_0\}$ in Y are statistically dependent and hence the MI-in-frequency obtained from (2.2) at these frequency pairs will be positive. Also the frequency components in Y at $\nu \in \{0, \pm 2\nu_0\}$ are dependent and hence the MI-in-frequency within Y at these frequencies will be positive.

2.3.1 Gaussian inputs to LTI filters

Let's consider the special case where $X(t)$, a Gaussian process with power spectral density $s_X(\nu)$ serves as the input to a linear, time-invariant (LTI) filter and $Y(t)$ is output observed in additive colored noise. The processes $X(t)$ and $Y(t)$ are related by

$$y(t) = h_1(t) * x(t) + h_2(t) * w(t), \tag{2.4}$$

where $*$ denotes convolution operation, $x(t), y(t)$ and $w(t)$ are sample paths of $X(t)$, $Y(t)$ and $W(t)$ respectively. W is a Gaussian process with power spectral density $s_W(\nu)$ and independent of X. $h_1(t)$ and $h_2(t)$ are continuous-time impulse responses of LTI filters, whose transfer functions are $H_1(j2\pi\nu)$ and $H_2(j2\pi\nu)$ respectively. Let

$\widetilde{X}(\nu)$, $\widetilde{W}(\nu)$ and $\widetilde{Y}(\nu)$ be the spectral processes of the Gaussian processes X, W and Y. Then from Theorem 3.2, we have

$$\begin{aligned}\left[d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\right] &\sim \mathcal{N}\left(\mathbf{0}, \tfrac{1}{2}s_X(\nu)\mathbf{I}\right),\\ \left[d\widetilde{W}_R(\nu), d\widetilde{W}_I(\nu)\right] &\sim \mathcal{N}\left(\mathbf{0}, \tfrac{1}{2}s_W(\nu)\mathbf{I}\right),\end{aligned} \tag{2.5}$$

where $\mathcal{N}(\mu, \Sigma)$ represents Gaussian distribution with mean μ and covariance Σ, $\mathbf{0}$ is the two element zero vector and $\mathbf{I}$ is the 2×2 identity matrix. In addition, we can show for this model in (2.4) that

$$d\widetilde{Y}(\nu) = H_1(j2\pi\nu)\, d\widetilde{X}(\nu) + H_2(j2\pi\nu)\, d\widetilde{W}(\nu). \tag{2.6}$$

The MI-in-frequency defined in (2.2) is further simplified for the model in (2.4) using (2.5), (2.6) and stated in the following theorem.

Theorem 3.1. *For the model given in* (2.4)*, the MI between* $X(t)$ *at frequency* ν_i *and* $Y(t)$ *at frequency* ν_j *is zero, when* $\nu_i \neq \nu_j$ *and the MI between* $X(t)$ *and* $Y(t)$ *at frequency* $\nu_i = \nu_j = \nu \neq 0$ *is*

$$\begin{aligned}\mathrm{MI}_{XY}(\nu,\nu) &= 2 \times \mathrm{I}\left(\left\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\right\}; d\widetilde{Y}_R(\nu)\right)\\ &= \log\left(1 + \tfrac{|H_1(j2\pi\nu)|^2 s_X(\nu)}{|H_2(j2\pi\nu)|^2 s_W(\nu)}\right).\end{aligned} \tag{2.7}$$

The proof of the above theorem is in the appendix. Note that at $\nu = 0$, the MI-in-frequency between X and Y is equal to $\mathrm{I}\left(\left\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\right\}; d\widetilde{Y}_R(\nu)\right)$, which is just half of the right hand side of (2.7). We intuitively expect different frequency components in the Gaussian input and its output from a linear system to be independent and Theorem 3.1 confirms that the proposed definition of MI-in-frequency agrees

with this intuition. In addition, the MI between X and Y is ∞ when $|H_2(j2\pi\nu)| = 0$, since the components of X and Y at such ν are linearly related. The MI between two different frequencies in $Y(t)$, generated from (2.4), is zero due to the linearity of the filters and Gaussian inputs. Furthermore, we can also show for the Gaussian processes X and Y related by (2.4) that MI-in-frequency is related to coherence $C_{XY}(\nu) \in [0,1]$, by $\mathrm{MI}_{XY}(\nu,\nu) = -\log(1 - C_{XY}(\nu))$. The proof is in the appendix. This result implies MI-in-frequency between Gaussian processes related by (2.4) can be estimated with the coherence. Theorem 3.1 also shows that MI-in-frequency between Gaussian processes related by (2.4) can be estimated by estimating the mutual information between $\left[d\widetilde{X}_R(\nu_i), d\widetilde{X}_I(\nu_i)\right]$ and $d\widetilde{Y}_R(\nu_j)$, a three dimensional estimate as opposed to the four dimensional estimation in general.

2.3.2 Discrete-time stochastic processes

We now extend the definition of MI-in-frequency between continuous-time stochastic processes in (2.2), (2.3) to discrete-time stochastic processes. In practice, we only have access to data samples from a discrete-time stochastic process, sampled at a given Nyquist sampling frequency F_s. Sampled signals have periodic spectra, with a period equalling F_s. In addition, components in the process with frequencies in the range $[F_s/2, F_s]$ correspond to negative frequencies [71]. Therefore, the actual frequency content in the signal is confined to $[0, F_s/2]$. We use normalized frequency $\lambda = \frac{\nu}{F_s} \in [0, 0.5]$ to describe the frequency axis in case of discrete-time stochastic processes, instead of ν which was used in case of continuous-time stochastic processes. The MI-in-frequency between discrete-time processes is therefore obtained by replacing ν_i, ν_j by the normalized frequencies $\lambda_1, \lambda_2 \in [0, 0.5]$ in (2.2), (2.3). Multivariate autoregressive (MVAR) models, commonly used to model electro-physiological signals

recorded from brain [61, 72], are a special case of the discrete-time equivalent of (2.4). The analytic expression for MI at frequency λ for MVAR models is therefore similarly obtained by replacing the frequencies ν by λ in (2.7), which is also equal to $-\log\left(1 - C_{XY}\left(\lambda\right)\right)$. This was independently suggested only for the special case of discrete-time Gaussian processes in [57, 73].

2.4 Data-Driven Estimation of MI-in-frequency

We describe two data-driven estimators–a kernel density based (KDMIF) and a nearest neighbor based (NNMIF) estimator to estimate MI-in-frequency, $\widehat{\mathrm{MI}}_{XY}\left(\lambda_i, \lambda_j\right)$, between λ_i component of X and λ_j component of Y. The input to both these algorithms are the N samples of X and Y. The first step in both KDMIF and NNMIF estimators involves estimating the samples of spectral process increments $d\widetilde{X}\left(\lambda_i\right)$ and $d\widetilde{Y}\left(\lambda_j\right)$, of X at λ_i and of Y at λ_j respectively. In the second step, the KDMIF estimator uses the kernel density based MI estimator [5, 58], whereas NNMIF estimator uses the k-nearest neighbor based MI estimator [58, 74] to estimate MI from the samples of spectral process increments, $d\widetilde{X}\left(\lambda_i\right)$ and $d\widetilde{Y}\left(\lambda_j\right)$.

2.4.1 Kernel Density Based MI-in-frequency (KDMIF) Estimator

Estimation of Samples of Spectral Process Increments

The first step of the algorithm is estimating the samples of spectral process increments of X and Y from N dependent data samples. We assume there is a finite memory in both these processes and chose an value for a parameter N_f, which encodes the length of dependence or memory in the data. We assume data in different windows are independent of each other. Ideally, consecutive windows should be separated to

ensure no dependence across windows and avoid the dependence across the window boundaries, but our simulation results demonstrate that not separating the windows doesn't affect performance significantly. In addition, N_f also determines the frequency resolution of our MI-in-frequency estimates. The N samples of X are split into N_s non-overlapping windows with $N_f = \frac{N}{N_s}$ data points in each window. Let us denote the samples in l^{th} window of X and Y respectively by two N_f element one-dimensional vectors, $\mathbf{x}^l$ and $\mathbf{y}^l$, for $l = 1, 2, \cdots, N_s$.

Let us now focus on estimating samples of the random variable $d\widetilde{X}(\lambda_i)$. Let $\mathcal{F}\left\{\mathbf{x}^l\right\}(\alpha)$ denote the discrete-time Fourier transform (DTFT) of $\mathbf{x}^l$ at normalized frequency α. For $\lambda_i = \frac{i}{N_f} \in [0,1]$ and $i \in [0, N_f - 1]$, let us define $d\widetilde{x}^l(\lambda_i)$ and integrated Fourier spectrum, $\widetilde{x}^l(\lambda_i)$, by

$$d\widetilde{x}^l(\lambda_i) = \mathcal{F}\left\{\mathbf{x}^l\right\}(\lambda_i) \text{ and } \widetilde{x}^l(\lambda_i) = \sum_{m=0}^{i} \mathcal{F}\left\{\mathbf{x}^l\right\}(\lambda_m). \tag{2.8}$$

It is stated in [66] that the random variable for which $\widetilde{x}^l(\lambda_i)$ is just one realization, tends to the spectral process of X at λ_i in mean of order γ, for any $\gamma > 0$, as the number of samples goes to infinity and assuming the underlying distribution is stationary and satisfies a mixing assumption. Also, $d\widetilde{x}^l(\lambda_i)$, which is the increment in $\widetilde{x}^l(\lambda_i)$ between λ_i and $\lambda_i + d\lambda$, is just the DTFT of the samples in window l. Calculating the DTFT with the FFT for each of the N_s windows separately yields an $N_f \times N_s$ matrix, whose i^{th} row, $\mathbf{d\widetilde{x}}(\lambda_i) = \left[d\widetilde{x}^1(\lambda_i), d\widetilde{x}^2(\lambda_i), \cdots, d\widetilde{x}^{N_s}(\lambda_i)\right]$ is the complex-valued vector containing N_s samples of $d\widetilde{X}(\lambda_i)$, the spectral process increments of X at $\lambda_i = \frac{i}{N_f}$. The l^{th} element of $\mathbf{d\widetilde{x}}(\lambda_i)$, $d\widetilde{x}^l(\lambda_i) = d\widetilde{x}^l_R(\lambda_i) + i d\widetilde{x}^l_I(\lambda_i)$, is a particular realization of $d\widetilde{X}(\lambda_i)$. A similar procedure is used to obtain the N_s samples of the spectral process increments of Y at $\lambda_j = \frac{j}{N_f}, j \in [0, N_f - 1]$ and the

resulting samples are denoted by $\mathbf{d\widetilde{y}}\left(\lambda_j\right) = \left[d\widetilde{y}^1\left(\lambda_j\right), d\widetilde{y}^2\left(\lambda_j\right), \cdots, d\widetilde{y}^{N_s}\left(\lambda_j\right)\right]$.

Estimating MI-in-frequency

The MI-in-frequency estimate is now obtained from the N_s samples, $\left(d\widetilde{x}_R^l(\lambda_i), d\widetilde{x}_I^l(\lambda_i)\right)$ and $\left(d\widetilde{y}_R^l(\lambda_j), d\widetilde{y}_I^l(\lambda_j)\right)$, for $l = 1, 2, \cdots, N_s$, using a kernel density based plug-in nonparametric estimator [58]. The N_s data samples are split into N_{tr} training and N_{ts} test samples. The training data is used to estimate the four-dimensional joint probability density $\mathrm{P}\left(d\widetilde{X}_R\left(\lambda_i\right), d\widetilde{X}_I\left(\lambda_i\right), d\widetilde{Y}_R\left(\lambda_j\right), d\widetilde{Y}_I\left(\lambda_j\right)\right)$. The density is estimated using a kernel density estimator with Gaussian kernels, the optimal bandwidth matrix selected using smoothed cross-validation criterion [5] and implemented using 'ks' package in R [75]. The joint density is marginalized to estimate the two-dimensional densities, $\mathrm{P}\left(d\widetilde{X}_R\left(\lambda_i\right), d\widetilde{X}_I\left(\lambda_i\right)\right)$ and $\mathrm{P}\left(d\widetilde{Y}_R\left(\lambda_j\right), d\widetilde{Y}_I\left(\lambda_j\right)\right)$, by recoginizing that the bandwidth matrix for the two-dimensional marginal is the appropriate 2×2 sub-matrix from the 4×4 bandwidth matrix for the joint density. The estimates of the joint and the marginal densities at the N_{ts} test samples are plugged into the following equation (2.9) to estimate MI-in-frequency.

$$\widehat{\mathrm{MI}}_{XY}\left(\lambda_i, \lambda_j\right) = \frac{1}{N_{ts}} \sum_l \log \frac{\widehat{\mathrm{P}}\left(d\widetilde{x}_R^l(\lambda_i), d\widetilde{x}_I^l(\lambda_i), d\widetilde{y}_R^l(\lambda_j), d\widetilde{y}_I^l(\lambda_j)\right)}{\widehat{\mathrm{P}}\left(d\widetilde{x}_R^l(\lambda_i), d\widetilde{x}_I^l(\lambda_i)\right)\widehat{\mathrm{P}}\left(d\widetilde{y}_R^l(\lambda_j), d\widetilde{y}_I^l(\lambda_j)\right)}. \tag{2.9}$$

2.4.2 Nearest Neighbor Based MI-in-frequency (NNMIF) Estimator

Estimation of Samples of Spectral Process Increments

The first step in the nearest neighbor based MI-in-frequency estimator is exactly same as that of KDMIF estimator. Following the steps described in section 2.4.1, we estimate $d\widetilde{x}^l\left(\lambda_i\right)$ and $d\widetilde{y}^l\left(\lambda_j\right)$, for $l = 1, 2, \cdots, N_s$, the N_s samples of the spectral process increments of X at λ_i and Y at λ_j respectively.

Estimating MI-in-frequency

$\mathrm{MI}_{XY}(\lambda_i, \lambda_j)$ is now estimated from $d\widetilde{x}^l(\lambda_i) \in \mathbb{R}^2$ and $d\widetilde{y}^l(\lambda_j) \in \mathbb{R}^2$, for $l = 1, 2, \cdots, N_s$ using nearest neighbor based MI estimator [74]. We apply the first version of the algorithm in [74] to two-dimensional random variables $d\widetilde{X}(\lambda_i)$ and $d\widetilde{Y}(\lambda_j)$ to compute $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$. Consider the joint four dimensional space $\left(d\widetilde{X}(\lambda_i), d\widetilde{Y}(\lambda_j)\right) \in \mathbb{R}^4$. The distance between two data points with indices $l_1, l_2 \in [1, N_s]$ is calculated using the infinity norm, according to $\max\left\{\|d\widetilde{x}^{l_1}(\lambda_i) - d\widetilde{x}^{l_2}(\lambda_i)\|, \|d\widetilde{y}^{l_1}(\lambda_j) - d\widetilde{y}^{l_2}(\lambda_j)\|\right\}$. Let ϵ_l denote the distance between the data sample $\left(d\widetilde{x}^l(\lambda_i), d\widetilde{y}^i(\lambda_j)\right)$ and its K^{th} nearest neighbor, for $l = 1, 2, \cdots, N_s$. We used $K = 3$ in this paper [67]. Let n_x^l and n_y^l denote the number of samples of $d\widetilde{X}(\lambda_i)$ and $d\widetilde{Y}(\lambda_j)$ within an infinity norm ball of radius less than ϵ_l centered at $d\widetilde{x}^l(\lambda_i)$ and $d\widetilde{y}^i(\lambda_j)$ respectively. From [74], the MI-in-frequency between X and Y at normalized frequencies λ_i and λ_j is given by

$$\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j) = \psi(K) + \psi(N_s) - \tfrac{1}{N_s}\sum_{l=1}^{N_s}\left(\psi\left(n_x^l + 1\right) + \psi\left(n_y^l + 1\right)\right), \tag{2.10}$$

where $\psi(\cdot)$ is the Digamma function.

2.4.3 Significance Testing

The statistical significance of the MI-in-frequency estimates obtained from both KD-MIF and NNMIF estimators is now tested using the following procedure. We permute the samples in the vector $\mathbf{d\widetilde{x}}(\lambda_i)$ randomly and estimate the MI-in-frequency between the permuted vector and the N_s samples of $d\widetilde{Y}(\lambda_j)$. Unlike just adding random phase or permuting the phase time series typically used to test the statistical significance of phase-amplitude coupling metrics [76], we permute the samples of spectral process increments since our metric can detect coupling across phase and amplitude jointly.

This process is repeated N_p times to obtain N_p permuted MI-in-frequency estimates, under the null hypothesis of independence. The permuted MI estimates will be almost zero, since the permutations make the spectral processes almost independent. If the actual MI estimate, $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$, is judged larger than all the permuted N_p estimates, then there is a statistically significant dependence between the processes at these frequencies.

2.5 MI between Data with Temporal Dependencies

We now use the MI-in-frequency to estimate mutual information between dependent data. The data-driven MI estimator, summarized in Algorithm 1, takes in N samples of X and Y as input and outputs the mutual information between X and Y, $\hat{\mathrm{I}}(X;Y)$, by estimating $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$, where $\lambda_i = \frac{i}{N_f}, \lambda_j = \frac{j}{N_f}, \forall (i,j)$ such that $i, j \in [0, N_f - 1]$.

2.5.1 Identifying Coupled Frequencies

The first step in our MI estimator involves estimating the MI-in-frequency, $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$, between $\lambda_i = \frac{i}{N_f}$ frequency component in X and $\lambda_j = \frac{j}{N_f}$ component in Y, for all (i,j) such that $i, j \in [0, N_f - 1]$ using either the KDMIF (section 2.4.1) or the NN-MIF (section 2.4.2) algorithms. Statistical significance of the resulting estimates is assessed using the procedure described in section 2.4.3. The resultant MI-in-frequency estimates across all frequency pairs can be graphically visualized by plotting the statistically significant MI-in-frequency estimates on a two-dimensional image grid, whose rows and columns correspond to frequencies of X and Y respectively. Let Λ_x and Λ_y respectively denote the set of frequency components of X and Y, such that for each $\lambda_{i_p} \in \Lambda_x$, there exists at least one $\lambda_{j_q} \in \Lambda_y$ for which $\widehat{\mathrm{MI}}_{XY}\left(\lambda_{i_p}, \lambda_{j_q}\right)$ is statistically

Algorithm 1: Mutual Information Estimator

Data: $(x[n], y[n])$, for $x[n], y[n] \in \mathbb{R}, n \in [0, N-1]$.

Result: $\hat{\mathrm{I}}(X;Y)$

Algorithm:

A) Estimate $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$ at all possible pairs (λ_i, λ_j), using either the KDMIF or the NNMIF estimator. Identify the sets Λ_x, Λ_y, containing frequency components of X, Y respectively with statistically significant MI-in-frequency estimates.

B) Let $d\widetilde{X}(\Lambda_x) = \left[d\widetilde{X}(\lambda_{j_1}), \cdots, d\widetilde{X}(\lambda_{j_P})\right] \in \mathbb{R}^{2P}$ and $d\widetilde{Y}(\Lambda_y) = \left[d\widetilde{Y}(\lambda_{l_1}), \cdots, d\widetilde{Y}\left(\lambda_{l_Q}\right)\right] \in \mathbb{R}^{2Q}$. The mutual information between X and Y is given by

$$\hat{\mathrm{I}}(X;Y) = \tfrac{1}{\max(P,Q)}\hat{\mathrm{I}}\left(d\widetilde{X}(\Lambda_x); d\widetilde{Y}(\Lambda_y)\right),$$

where the right hand side is estimated from N_s i.i.d. samples using any nonparametric MI estimator [58].

significant and vice-versa.

2.5.2 Estimating Mutual Information

The final step in our algorithm estimates MI between the spectral process increments of X and Y at frequencies in Λ_x and Λ_y respectively. With P, Q denoting the cardinality of Λ_x, Λ_y respectively, let $d\widetilde{X}(\Lambda_x)$ and $d\widetilde{Y}(\Lambda_y)$ denote the $2P$ and $2Q$-dimensional random vector comprising the spectral process increments of X, Y at all frequencies in Λ_x and Λ_y respectively. We already computed N_s i.i.d. samples of these two random vectors to estimate MI-in-frequency estimates in the previous step of this algorithm. The desired MI estimate is computed from the mutual information between $d\widetilde{X}(\Lambda_x)$ and $d\widetilde{Y}(\Lambda_y)$, which is estimated using the k-nearest neighbor based

estimator developed in [74], according to

$$\hat{\mathrm{I}}\left(X;Y\right)=\frac{1}{\max\left(P,Q\right)}\hat{\mathrm{I}}\left(d\widetilde{X}\left(\Lambda_x\right);d\widetilde{Y}\left(\Lambda_y\right)\right). \tag{2.11}$$

The MI estimator in (2.11) can be further simplified for discrete-time Gaussian processes. Without loss of generality, consider two Gaussian processes X and Y, related by

$$y[n]=h_1[n]*x[n]+h_2[n]*w[n], \tag{2.12}$$

where $h_1[n], h_2[n]$ are linear time-invariant (LTI) filters and W is white Gaussian noise independent of X. For the model in (2.12), which is the discrete-time equivalent of (2.4), the data-driven estimation in (2.11) can be further simplified to

$$\hat{\mathrm{I}}\left(X;Y\right)=\tfrac{1}{N_f}\sum_{i=0}^{N_f/2}\widehat{\mathrm{MI}}_{XY}\left(\lambda_i;\lambda_i\right),\ \text{where}\ \lambda_i=\tfrac{i}{N_f}. \tag{2.13}$$

This result is obtained because linear models do not introduce cross-frequency dependencies and because negative frequencies do not carry any extra information. Furthermore, the relationship between the MI and their MI-in-frequency for two processes related by (2.12) is stated in the following theorem.

Theorem 5.1. *Consider two discrete-time Gaussian stochastic processes X and Y related by of* (2.12)*. The mutual information between these processes, a scalar, is given by*

$$\mathrm{I}\left(X;Y\right)=\int_0^{0.5}\mathrm{MI}_{XY}\left(\lambda,\lambda\right)d\lambda. \tag{2.14}$$

The proof of the above theorem is in the appendix. This theorem means that MI between two Gaussian processes over the entire time can be obtained by integrating the contribution from each frequency component. It is easy to see that the right hand side of (2.13) is just the Riemann sum of the integral on the right hand side of (2.14), which converges to the true value as N_f tends to infinity. This implies our estimator converges to the true value for discrete-time Gaussian processes.

Note that the MI estimation algorithm does not make any parametric assumptions on the underlying model between X and Y. The computation of MI via (2.11) can be greatly simplified by clustering the frequencies in Λ_x and Λ_y into groups without any significant dependencies across groups and using the chain rule of mutual information. In addition, if we observe after the first step that significant MI-in-frequency estimates occur only at $(\lambda_i, \lambda_i), \forall i \in \left[0, N_f - 1\right]$, then the MI can be estimated using (2.13).

2.6 Performance Evaluation on Simulated Data

The performance of the data-driven MI-in-frequency and mutual information estimators described in section 2.4 and section 2.5 respectively is validated on simulated data. The statistical significance of the estimates was assessed using the procedure described in section 2.4.3. In addition, we compare the performance of the MI-in-frequency estimators against modulation index [59, 60, 62], a commonly used phase-amplitude coupling metric in neuroscience.

2.6.1 Comparing the KDMIF and NNMIF Estimators

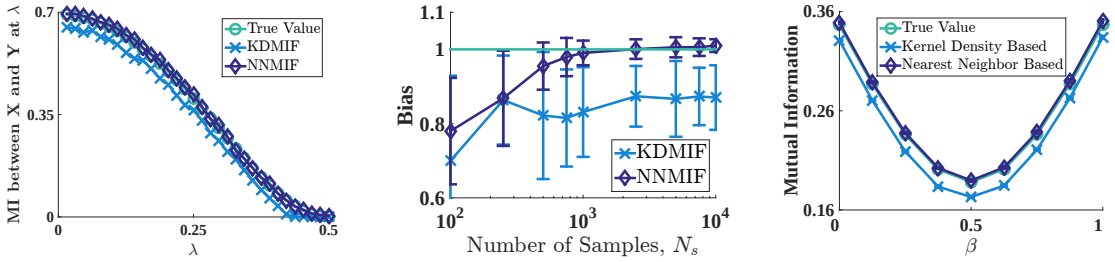
Consider two stochastic processes X and Y, where X is a white Gaussian process with standard deviation σ_x and Y is obtained by

$$y[n] = h[n] * x[n] + w[n], \tag{2.15}$$

where W is a white Gaussian process with standard deviation σ_w that is independent of X and $h[n]$ is a linear time-invariant filter. We compared the performance of the kernel density based and nearest neighbor based estimators by benchmarking the estimates against the true value of MI-in-frequency and the mutual information between X and Y for the model in (2.15). We used two different filers: a two-tap low pass filter, $h[n] = [\beta, 1-\beta]$, for $\beta \in [0, 1]$ and a 33-tap bandpass filter with passband in $[0.15, 0.35]$ normalized frequency range. We observed that modulation index, a popular CFC metric, was unable to correctly detect and quantify the strength of cross-frequency coupling for both these models.

Lowpass Filter

The samples of X and Y are generated from (2.15) with $\sigma_x = \sigma_w = 1$ and a lowpass filter with unit-impulse response $[\beta, 1-\beta]$, for various values of $\beta \in [0, 1]$. The true value of MI-in-frequency at normalized frequency $\lambda \in [0, 0.5]$ is obtained substituting the parameters of this model in (2.7) and is plotted in Fig. 2.1a for $\beta = 0.5$. In addition, the MI-in-frequency estimated by the KDMIF and NNMIF algorithms from $N = 64 \times 10^4$ data samples, with $N_f = 64, N_s = 10^4$ is also plotted in Fig. 2.1a. It is seen that the estimates from both algorithms follow the true value closely, without the knowledge of the underlying model. In addition, we evaluate the bias and the

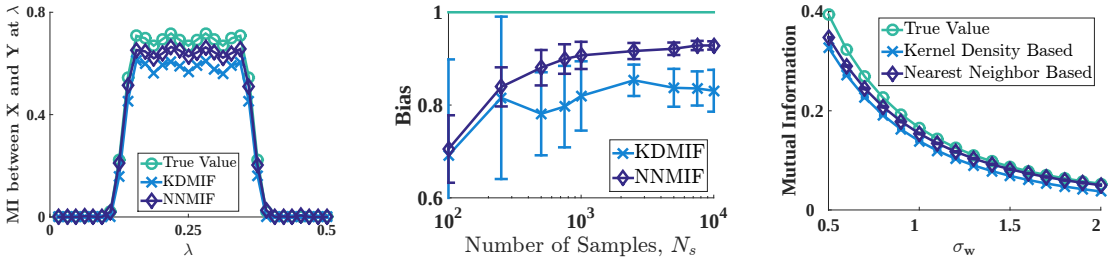


(a) MI-in-frequency (b) Bias of the Estimators (c) Mutual Information

Figure 2.1 : Comparing the performance of the kernel density based and nearest neighbor based estimators, KDMIF and NNMIF respectively, on simulated generated from (2.15) using a two-tap lowpass filter. In Fig. 2.1a, the MI-in-frequency estimates obtained from KDMIF and NNMIF estimators along with the true value of MI-in-frequency are plotted against the normalized frequency λ for $\beta = 0.5$. Fig. 2.1b plots the bias (mean of the ratio of the estimate and the true value in the filter passband) against the number of data samples used for estimation for $\beta = 0.5$. Fig. 2.1c plots the MI estimate between X and Y obtained from kernel density and nearest neighbor algorithms along with the true value of MI for $\beta \in [0, 1]$.

rate of convergence of both these algorithms as a function of N_s, with $N_f = 64$ in Fig. 2.1b. The bias is defined as the average value of the ratio of MI-in-frequency estimate and its true value in the passband of the lowpass filter. We observe that the NNMIF algorithm converges faster and has lower bias than the KDMIF algorithm. We now use both these algorithms to estimate the mutual information between X and Y for $\beta \in [0, 1]$. The analytical expression for the true value of MI* for this model is derived in [20]. It is evident from Fig. 2.1c that the MI estimates obtained from the nearest neighbor estimator are closer to the true value than those from the kernel density estimator.

*Note that for this particular model, mutual information is equal to the directed information from X to Y and the analytical expression is given in equation (18) in [20].

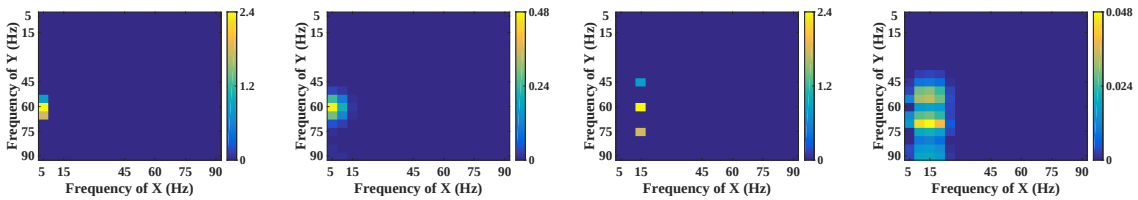


(a) MI-in-frequency (b) Bias of the Estimators (c) Mutual Information

Figure 2.2 : Comparing the performance of the kernel density based and nearest neighbor based estimators, KDMIF and NNMIF respectively, on simulated generated from (2.15) using a 33-tap bandpass filter with passband in $[0.15, 0.35]$ normalized frequency. In Fig. 2.2a, the MI-in-frequency estimates obtained from KDMIF and NNMIF estimators along with the true value of MI-in-frequency are plotted against the normalized frequency λ for $\sigma_w = 1$. Fig. 2.2b plots the bias (mean of the ratio of the estimate and the true value in the filter passband) against the number of data samples used for estimation for $\sigma_w = 1$. Fig. 2.2c plots the plots the MI estimate between X and Y from kernel density and nearest neighbor algorithms along with the true value of MI for different values of $\sigma_w \in [0.5, 2]$.

Bandpass Filter

The samples of X are generated from a standard white Gaussian random process with $\sigma_x = 1$ and those of Y are generated from (2.15) using a 33-tap finite-impulse response bandpass filter with passband in $[0.15, 0.35]$ normalized frequency range for different values of noise standard deviation, $\sigma_w \in [0.5, 2]$. We used the kernel density and the nearest neighbor based algorithms to estimate the MI-in-frequency and the mutual information between X and Y. The true value of MI-in-frequency is obtained from (2.7) and of mutual information is numerically calculated using power spectral density (chapter 10 in [56]). It is clear from Fig. 2.2b that the nearest neighbor based algorithm converges to the true value faster than the kernel density based algorithm. The nearest neighbor based algorithms also provides more accurate estimates of both MI-in-frequency and mutual information between X and Y, as evident from Fig. 2.2a,



(a) MI-in-frequency (b) Modulation Index (c) MI-in-frequency (d) Modulation Index

Figure 2.3 : Comparing the performance of MI-in-frequency against modulation index in detecting cross-frequency coupling in data generated from (2.16). In Fig. 2.3a and Fig. 2.3b, MI-in-frequency estimates obtained from nearest neighbor algorithm and modulation index are plotted respectively, when $f_l = 5$ Hz and $f_h = 60$ Hz in (2.16). Fig. 2.3c and Fig. 2.3d respectively plot the MI-in-frequency estimates and modulation index estimates, when $f_l = 15$ Hz and $f_h = 60$ Hz in (2.16).

Fig 2.2c respectively. In addition, nearest neighbor based MI-in-frequency algorithm runs faster than kernel density based algorithm. We therefore, conclude that the nearest neighbor based MI-in-frequency algorithm outperforms kernel density based algorithms and only depict the results obtained from nearest neighbor based algorithm in the remainder of the paper.

2.6.2 Comparison with Modulation Index

We now compare the effectiveness of MI-in-frequency against modulation index in detecting cross-frequency coupling, using the simulated model commonly used to validate CFC metrics [62, 69, 77]. Consider two random cosine waves, $s_l[n]$ and $s_h[n]$, at frequencies f_l and f_h respectively. Let f_s denote the sampling frequency. The samples of time-series X and Y are generated from the following model:

$$s_l[n] = A\cos\left(2\pi\tfrac{f_l}{f_s}n + \theta\right), s_h[n] = A\cos\left(2\pi\tfrac{f_h}{f_s}n + \theta\right)$$
$$x[n] = s_l[n] + w_1[n], y[n] = (1 + s_l[n])\, s_h[n] + w_2[n], \tag{2.16}$$

where A is a Rayleigh random variable with parameter 1 and θ is a uniformly distributed random variable between 0 and 2π that is independent of A. $w_1[n], w_2[n]$ are samples of i.i.d white Gaussian noise process with standard deviation 1. We generated samples from this model with $f_l = 5$ Hz, $f_h = 60$ Hz and $f_s = 200$ Hz. MI-in-frequency between X and Y is estimated using the nearest neighbor based algorithm from $N = 40 \times 10^4$ samples with $N_s = 10^4$ and plotted in Fig. 2.3a. Modulation index between X and Y estimated by using the Matlab toolbox [69], with the amplitude envelope estimated using the Hilbert transform and is plotted in Fig. 2.3b. It is clear that both MI-in-frequency and modulation index successfully detect the cross-frequency coupling between 5 Hz component of X and $\{55, 60, 65\}$ Hz components of Y for these parameter values. We then generated X and Y from (2.16) with $f_l = 15$ Hz and all other parameter values unchanged. Fig. 2.3c plots the MI-in-frequency estimates obtained via NNMIF algorithm and as expected, we detect the CFC between 15 Hz component of X and $\{45, 60, 75\}$ Hz components of Y. However, modulation index, depicted in Fig. 2.3d, was not able to correctly detect the CFC between X and Y for these parameter values. In addition, the strength of the modulation index decreased from around 0.4 when $f_l = 5$ Hz in Fig. 2.3b to 0.04 when $f_l = 15$ Hz in Fig. 2.3d. This is because metrics like modulation index can only detect the CFC correctly with good frequency resolution only when one of the frequencies involved is very small compared to the other frequency. Otherwise, the bandwidth of the filter used to extract the phase and the amplitude envelope should be larger, which will reduce the frequency resolution in the estimated CFC (note the smearing in Fig. 2.3d, when compared to Fig. 2.3b) [62,77]. In addition, we tested modulation index on data generated from (2.15) and (2.17) and found that modulation index is unable to detect the cross-frequency coupling for these relationships. This is not surprising since the

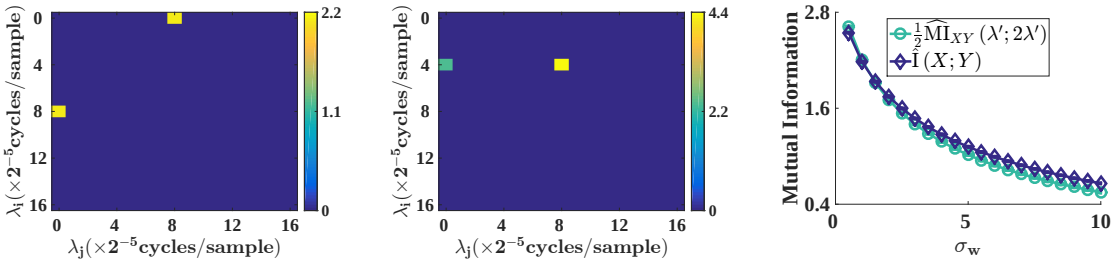
modulation index like metrics are tuned to detect CFC when the underlying coupling is of the form in (2.16), whereas the MI-in-frequency defined in this paper overcomes this shortcoming, which is evident from its performance on various simulated models.

2.6.3 Nonlinear Models

We now consider square nonlinearity, where the random processes X and Y are related by

$$y[n] = x[n]^2 + w[n], \tag{2.17}$$

where $w[n[$ is white Gaussian noise with standard deviation σ_w. Modulation index was not able to detect and quantify the cross-frequency coupling for this model. We estimated the MI-in-frequency between frequency components within Y, $\widehat{\mathrm{MI}}_{YY}(\lambda_i, \lambda_j)$, between the frequency components of X and Y, $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$, and the mutual information between X and Y, $\hat{\mathrm{I}}(X;Y)$, from $N = 32 \times 10^4$ samples of X and Y with $N_s = 10^4$, for different values of noise standard deviation, $\sigma_w \in [0.5, 2]$. Computing the true value of MI-in-frequency and mutual information is nontrivial because of the nonlinearity. The performance of the algorithms is assessed by checking if they detect the cross-frequency coupling at expected frequency pairs and by checking if the mutual information estimates decrease with increasing noise power as expected. We considered two different models for the stochastic process X, such that its samples are dependent across time.



(a) $\widehat{\mathrm{MI}}_{YY}(\lambda_i, \lambda_j)$ (b) $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$ (c) MI between X and Y

Figure 2.4 : (a) MI-in-frequency estimates from the nearest neighbor based algorithm between the frequency components within the random processes Y, obtained from the single cosine data-generation model, (2.18) with $\sigma_w = 1$. (b) MI-in-frequency estimates between random processes X and Y related by the single cosine data-generation model with $\sigma_w = 1$. It is clear that MI-in-frequency estimator correctly identifies the pairwise frequency dependencies. (c) MI-in-frequency between X at λ_0 and Y at $2\lambda_0$, $\widehat{\mathrm{MI}}_{XY\lambda_2}(\lambda_0, 2\lambda_0)$, obtained from (2.10) along with the MI estimate between X and Y, $\hat{\mathrm{I}}(X;Y)$, obtained from Algorithm 1 for various values of the noise standard deviation, σ_w.

Random Cosine with Squared Nonlinearity

The samples of X are generated from a random cosine wave,

$$x[n] = A\cos(2\pi\lambda_0 n + \theta), \tag{2.18}$$

where A is a Rayleigh random variable with parameter 1, θ is a uniform random variable between 0 and 2π that is independent of A and $\lambda_0 = \frac{4}{32}$. It is easy to see that frequency components of X are statistically independent and this is confirmed by the NNMIF estimator. However, because of the square nonlinearity in (2.17), the DC component of Y and the $2\lambda_0$ component of Y will be statistically dependent and this is confirmed by Fig. 2.4a, which plots the MI-in-frequency between components of Y generated with $\sigma_w = 1$ using the NNMIF algorithm. In addition, the common information between these two processes will be present between λ_0 component

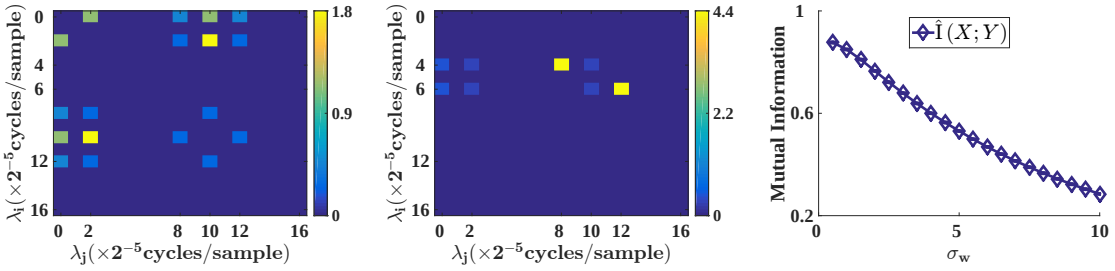
of X and the $\{0, 2\lambda_0\}$ components of Y. This cross-frequency dependence is confirmed by Fig. 2.4b, which plots the estimates of MI-in-frequency between X and Y obtained by the NNMIF algorithm from (2.10): we observe that significant dependencies occur only at $(\lambda_0, 0)$ and $(\lambda_0, 2\lambda_0)$ frequency pairs. As a result, $P = 1, Q = 2$. The MI estimate from Algorithm 1, $\hat{\mathrm{I}}(X;Y) = \frac{1}{2}\hat{\mathrm{I}}\left(d\widetilde{X}(\lambda_0); \left\{d\widetilde{Y}(0), d\widetilde{Y}(2\lambda_0)\right\}\right)$ is plotted in Fig. 2.4b. The MI estimate decreases with increasing σ_w as expected. In addition, we note for this model that the DC component of Y does not contain any extra information about X, given the $2\lambda_0$ component of Y. Therefore, we expect $\frac{1}{2}\hat{\mathrm{I}}\left(d\widetilde{X}(\lambda_0); \left\{d\widetilde{Y}(0), d\widetilde{Y}(2\lambda_0)\right\}\right) = \frac{1}{2}\widehat{\mathrm{MI}}_{XY}(\lambda_0; 2\lambda_0)$, a result verified in Fig. 2.4c, since the two curves are very close.

Two Random Cosines with Squared Nonlinearity

The samples of random process X are generated according to

$$x[n] = A_1 \cos(2\pi\lambda_1 n + \theta_1) + A_2 \cos(2\pi\lambda_2 n + \theta_2), \tag{2.19}$$

where A_1, A_2 are independent Rayleigh random variables with parameter 1, θ_1, θ_2 are independent uniformly distributed random variables between 0 and 2π that are independent of A_1, A_2, and $\lambda_1 = \frac{4}{32}, \lambda_2 = \frac{6}{32}$. As before, the frequency components of X are statistically independent. However, after some basic algebra, it is easy to see that the all possible pairs of frequency components of Y in $\{0, \lambda_2 - \lambda_1, 2\lambda_1, \lambda_2 + \lambda_1, 2\lambda_2\}$ are statistically dependent, except for $(2\lambda_1, 2\lambda_2)$ frequency pair, and we expect to see statistically significant MI-in-frequency estimates between these frequency components. This is confirmed by Fig. 2.5a, which plots the MI-in-frequency estimates within Y, generated with $\sigma_w = 1$ and obtained by the NNMIF algorithm. In addition,



(a) $\widehat{\mathrm{MI}}_{YY}(\lambda_i, \lambda_j)$ (b) $\widehat{\mathrm{MI}}_{XY}(\lambda_i, \lambda_j)$ (c) MI between X and Y

Figure 2.5 : (a) MI-in-frequency estimates from the nearest neighbor based algorithm between the frequency components within the random processes Y, obtained from the two cosine data-generation model, (2.19). (b) MI-in-frequency estimates between random processes X and Y related by the two cosine data-generation model. It is clear that MI-in-frequency estimator correctly identifies the pairwise frequency dependencies between X and Y. (c) $\hat{\mathrm{I}}(X;Y)$, the MI estimate between X and Y obtained from Algorithm 1 for various values of the noise standard deviation, σ_w.

the pairwise frequency dependencies between X and Y occur at $(\lambda_1, 0)$, $(\lambda_1, \lambda_2 - \lambda_1)$, $(\lambda_1, 2\lambda_1)$, $(\lambda_1, \lambda_2 + \lambda_1)$, $(\lambda_2, 0)$, $(\lambda_2, \lambda_2 - \lambda_1)$, $(\lambda_2, \lambda_2 + \lambda_1)$ and $(\lambda_2, 2\lambda_2)$. Fig. 2.5b plots the estimates of pairwise MI-in-frequency between X and Y generated with $\sigma_w = 1$ and obtained by the data-driven NNMIF algorithm using (2.10). The algorithm correctly identifies all the dependent frequency pairs and $P = 2, Q = 5$. We then apply the algorithm described in section 2.5 and plot the estimates the MI for different values of noise standard deviation σ_w in Fig. 2.5c. Again, the MI decreases with increasing noise power, as expected. These different models validate the superiority of MI-in-frequency over other existing metrics to detect cross-frequency coupling and also demonstrate the performance and accuracy of the data-driven MI-in-frequency and MI estimators.

2.7 Conclusions

Motivated to understand frequency coupling in electrophysiological recordings from brain, we defined MI-in-frequency between stochastic processes, that are not necessarily Gaussian and estimated it using data-driven estimators. We compared the performance of kernel density based and nearest neighbor based MI-in-frequency estimator. Unlike in the problem of MI estimation from i.i.d. data [67], where kernel density estimator was superior for very short datasets with high noise levels and nearest neighbor based estimator was superior for short data datasets or the problem using MI for feature selection, where kernel density estimation was superior [68], we found that for the problem of estimating MI-in-frequency, the nearest neighbor based estimator outperforms the kernel based estimator in accuracy, convergence and computational complexity.

We also compared the performance of MI-in-frequency against modulation index, a popular phase-amplitude coupling metric. The main advantages of the MI-in-frequency approach over existing methods to estimate CFC is that it detects statistical independence, detects dependencies across phase and amplitude, and does not dependent on parameters like the filter bandwidth. Our approach will need more data when compared with the other approaches like coherence, since MI-in-frequency detects statistical independence. From the simulation results on linear models, we need about 10^3 samples to be within 10% of the true value. For the ECoG data sampled at 1 KHz and a desired spectral resolution of 10 Hz, this implies the total number of data samples is of the order of 100 seconds or couple of minutes, which is roughly the size of preictal, ictal and postictal windows used in chapter 4. Also, it is reported in [76] that the minimum size of window required to estimate phase-amplitude coupling is 10 seconds. The time-frequency resolution of our Fourier based

approach can be improved by moving to wavelet based analysis in the future. We also assume the data is stationary in each observation window, which is not necessarily accurate. We can potentially relax this assumption by utilizing time-frequency distributions and develop some heuristics, however, the inherent trade-off involved is that we are not detecting statistical independence anymore. It is also straightforward to define and estimate conditional MI-in-frequency to eliminate indirect coupling estimated between two signals because of a third signal which is coupled to both. We also utilized the MI-in-frequency to estimate mutual information between dependent data. In summary, we developed a first of its kind metric to detect statistical independence in frequency and for the first time, utilize frequency domain to estimate mutual information over time.

Chapter 3

Directed Information

3.1 Introduction

We describe the proposed model-based and data-driven estimators of directed information, a versatile metric that can detect and quantify the causal connectivity between the data recorded from multiple sensors, in this chapter. Effective or causal connectivity [41] from one data stream X to another data stream Y is the reduction in uncertainty in predicting the future values of Y, given the causal past of X. The sensors correspond to the ECoG electrodes in case of recordings from epilepsy patients and we are interested in inferring the causal connectivity between the brain regions.

Estimating causal connectivity from electrophysiological recordings of brain has been the focus of many papers. A good summary is provided in [50]. The causality referred to in this thesis is in the Wiener-Granger causal sense [78]. Metrics based on Granger causality (GC) [52, 79] and information theory like transfer entropy [80] are commonly used to estimate causal connectivity between continuous-valued data. However, these techniques are well-suited only for a specific model and subset of the recorded signals. For instance, GC-based measures are applicable only for data from multivariate autoregressive (MVAR) processes. Continuous-valued recordings from brain like ECoG data, are often modeled using linear MVAR model [50, 52], even though associations between ECoG recordings are likely nonlinear [72, 81].

We propose to develop a causal metric that would be applicable to diverse models and especially for different types of electrophysiological recordings of brain. Directed information, used to infer causal connections between spike trains in [82–84], can indeed be further developed into a general technique to estimate causal connectivity. The definition of DI is based on the underlying probability distribution and no assumptions are imposed on the underlying distributions. Directed information was developed for discrete-valued time-series in [85–87] and nonparametrically estimated in [88]. DI quantifies the amount of causal information about one time-series that is explained by the other time-series [82]. Modified time-lagged directed information is proposed in [89, 90] to reduce the computational complexity of estimating directed information. DI is also used in many other applications [91–93]. The definition of DI is broadened to the class of continuous-valued processes like ECoG signals in this chapter. [18,89,90,94]. If the data is assumed to be from a MVAR process with Gaussian white noise, DI is equivalent to Granger causality [94] and if the data satisfies the general Markov condition, DI is very closely related to transfer entropy [89, 90]. The main advantage of DI over other existing techniques in neuroscience is that DI is applicable to a large class of electrophysiological recordings from brain, including spike trains, EEG and ECoG, and is not restricted to a particular class of models.

We developed an almost surely convergent model-based and data-driven estimators of DI, inspired by prior work [82, 95]. The performance of the proposed DI estimators was validated on linear and nonlinear simulated models and compared with the Granger causality metric [79, 96]. The statistical significance of the causal connection inferred using DI and GC estimates was demonstrated using an adaptation [97] of stationary bootstrap [98]. The main algorithmic contributions of this chapter is in developing an almost surely convergent model-based and data-driven DI

estimator, described in sections 3.3 and 3.4.

3.2 Directed Information

Consider the N samples recorded at a sampling frequency F_s from each sensor, that is recording data from the system under study. Without loss of generality, let us focus on data from two sensors (or equivalently channels), X and Y. The N samples recorded from the two channels are denoted by $X^N = (x[1], x[2], \cdots, x[N])^{\mathrm{T}}$ and $Y^N = (y[1], y[2], \cdots, y[N])^{\mathrm{T}}$. Also let W denote the matrix of samples recorded from a group of channels excluding X and Y. For notational simplicity, the elements corresponding to the non-positive subscripts are treated as empty sets and the subscripts are not shown when equal to 1.

The directed information, $I\left(X^N \to Y^N\right)$, from N samples of continuous-valued random process X to those of Y is defined as

$$I\left(X^N \to Y^N\right) = h\left(Y^N\right) - h\left(Y^N \| X^N\right), \tag{3.1}$$

where $h\left(Y^N\right)$ is the differential entropy of the N-dimensional continuous random vector Y^N [95] and $h\left(Y^N \| X^N\right)$ is the causally conditioned differential entropy of Y^N causally conditioned on X^N. The causally conditioned differential entropy is defined as

$$h\left(Y^N \| X^N\right) = \sum_{n=1}^{N} h\left(y[n] | Y^{n-1}, X^n\right). \tag{3.2}$$

The definitions of DI and causally conditioned differential entropy in (3.1) and (3.2) are obtained by broadening the definitions of the same quantities from discrete-time, discrete-valued random processes [86, 87] to discrete-time, continuous-valued processes [19]. One of the main differences between discrete-valued and continuous-

valued random processes is that the entropy of a discrete-valued process is always non-negative, whereas the differential entropy of a continuous-valued process can be negative [95]. However, DI is always non-negative since conditioning cannot increase differential entropy [95], i.e., $I\left(X^N \to Y^N\right) \geq 0$. DI can be interpreted as the number of bits of uncertainty in one process that is causally explained away by the other process. If $I\left(X^N \to Y^N\right) = 0$, then there is no causal influence from X to Y. The DI is not a symmetric metric in general, i.e., $I\left(X^N \to Y^N\right) \neq I\left(Y^N \to X^N\right)$. Note that DI can also be expressed in terms of conditional mutual information [95], $I\left(y[n]; X^n | Y^{n-1}\right)$, as

$$\begin{aligned} I\left(X^N \to Y^N\right) &= \sum_{n=1}^{N} \left\{h\left(y[n]|Y^{n-1}\right) - h\left(y[n]|Y^{n-1}, X^n\right)\right\} \\ &= \sum_{n=1}^{N} I\left(y[n]; X^n|Y^{n-1}\right). \end{aligned} \tag{3.3}$$

Now, the DI between the time-series X and Y is defined as

$$\begin{aligned} I\big(X \to Y\big) &= \lim_{N\to\infty} \tfrac{1}{N} I\big(X^N \to Y^N\big) \\ &= \lim_{N\to\infty} \tfrac{1}{N} h\big(Y^N\big) - \lim_{N\to\infty} \tfrac{1}{N} h\big(Y^N \| X^N\big) \\ &= h\big(Y\big) - h\big(Y\|X\big), \end{aligned} \tag{3.4}$$

provided the limits exist. $h\left(Y\right)$ and $h\left(Y\|X\right)$ are respectively the differential entropy of Y and the causally conditioned differential entropy of Y given X. The DI from Y to X is also similarly defined.

Furthermore, the DI defined earlier is easily extended to define directed information from X to Y, causally conditioned on W. Note that W comprises the samples recorded from a group of sensors (or equivalently channels) excluding X and Y. The

causally conditioned DI, $I\left(X \to Y \| W\right)$, is defined as

$$\begin{aligned} I\left(X \to Y \parallel W\right) &= \lim_{N\to\infty} \tfrac{1}{N} I\left(X^N \to Y^N \| W^N\right) \\ &= \lim_{N\to\infty} \tfrac{1}{N} \left\{ h\left(Y^N \| W^N\right) - h\left(Y^N \| X^N, W^N\right) \right\}, \\ &= h\left(Y \| W\right) - h\left(Y \| W, X\right), \end{aligned} \tag{3.5}$$

where $h\left(Y^N \| X^N, W^N\right) = \sum_{n=1}^{N} h\big(y[n] | Y^{n-1}, X^n, W^n\big)$ is the differential entropy of Y^N causally conditioned on X^N and W^N, $h\left(Y \| W\right)$ is the causal conditioned differential entropy of Y given W and $h\left(Y \| W, X\right)$ is the causally conditioned differential entropy of Y given the causal past of W and X. We use the directed information defined here to learn the causal connectivity graph between all ECoG channels and identify the SOZ of epileptic patients in chapter 4.

3.3 Universal Estimator for Directed Information

A universal estimator for directed information between channels X and Y, $I\left(X \to Y\right)$, and the causally conditioned DI, $I\left(X \to Y \| W\right)$ is developed in this section. The proposed estimator is universal and is shown to be almost surely convergent assuming that the causal conditional likelihood (CCL) is known. If CCL is not known and is estimated, then the convergence of the proposed DI estimator is dependent on the CCL estimator. The ideas used in developing the proposed DI estimator are inspired by prior work [82,95]. Without loss of generality, we will first focus on estimating the pairwise DI, $I\left(X \to Y\right)$. We will then outline the procedure to extend this pairwise DI estimator to estimate the causally conditioned DI, $I\left(X \to Y \| W\right)$. The inputs to the proposed pairwise DI estimator are the observed N samples of time-series X and Y. The main idea is to develop an almost surely convergent estimator for

the entropies in (3.4) and the difference between the two entropy estimates is an almost surely convergent estimate for $I(X \to Y)$. Let us first focus on the causally conditioned differential entropy estimator, $\hat{h}(Y\|X)$.

Assumption 1 - The random processes X and Y are assumed to be stationary, ergodic and Markovian in the observed time-window. These are reasonable assumptions to model ECoG data. First, an implicit assumption in the problem of estimating the causal connectivity from a ECoG data segment is that the causal connectivity does not vary in this segment, which is mathematically captured by stationarity. The entire ECoG data record is usually not stationary and stationary segments are identified using either sliding windows [99] or change-point detection algorithms [9]. We used the sliding window approach in this thesis. A crucial parameter in this process is the length of the window in which data is assumed to be stationary. It is also important to realize that we need a minimum amount of data points to reliably estimate any unknown parameters involved. It is recommended that the number of data points should be much larger (as a thumb rule, at least an order of magnitude larger) than the number of parameters to be estimated [99]. Directed information is then estimated in each stationary segment using the algorithm proposed in this section. Directed information for the entire time-series is the sum of the DI estimates from each stationary segment and is interpreted as the total amount of uncertainty in one time-series in the entire recording window that is explained by the other time-series. Second, ergodicity is required to ensure that the estimates from long-enough recording windows converge to the true value. Finally, the Markovian assumption captures the dependence of the current activity on the past activity at different electrodes. Let the current sample of the time-series Y depend on the past J_{yy} and past K_{yx} samples of the time-series Y and X respectively. Note that (J_{yy}, K_{yx}) are unknown

and should be estimated from data. The explicit model of the dependence is captured by the causal likelihood of y_n conditioned on the past activity at electrodes X and Y. This CCL is denoted by $\mathrm{P}\big(y[n]|Y_{n-J_{yy}}^{n-1}, X_{n-K_{yx}+1}^{n}\big)$ and can be estimated using either a model-based or a data-driven approach. Let us assume for now that CCL is known.

Assumption 2 - Let us also assume that differential entropy of the first sample, $y[1]$, of time-series Y exists and that for some time-index $l \in [1, N]$, the conditional differential entropy of $y[l]$, conditioned on $Y_{l-J_{yy}}^{l-1}$ and $X_{l-K_{yx}+1}^{l}$ also exists, i.e., $h\left(y[1]\right), h\left(y[l]|Y_{l-J_{yy}}^{l-1}, X_{l-K_{yx}+1}^{l}\right) \in \mathbb{R}$.

Lemma 3.1. *Let Assumptions 1 and 2 hold. Then* $h\left(Y\right),\ h\left(Y\|X\right),\ I\left(X \to Y\right)$ *exists and are in* $\mathbb{R}$.

Proof. Stationarity and the property that conditioning cannot increase the differential entropy are the main ideas in the proof, which is in the Appendix C.1. □

Lemma 3.2. *Let Assumptions 1 and 2 hold. Then for some time-index* l

$$\tfrac{1}{N} h\left(Y^N \| X^N\right) = \mathbb{E}\left[-\log \mathrm{P}\left(y[l]|Y_{l-J_{yy}}^{l-1}, X_{l-(K_{yx}-1)}^{l}\right)\right]. \tag{3.6}$$

Proof. The proof uses the definition of causally conditioned differential entropy (3.2), the Markovian and the stationarity assumptions. The proof is in the Appendix C.1. □

Theorem 3.1. *Let Assumptions 1 and 2 hold. Then the almost surely convergent causally conditioned differential entropy estimator is*

$$\hat{h}\left(Y\|X\right) = \tfrac{1}{N} \sum_{n=1}^{N} \left\{-\log \mathrm{P}\left(y[n]|Y_{n-J_{yy}}^{n-1}, X_{n-(K_{yx}-1)}^{n}\right)\right\}. \tag{3.7}$$

Proof. The proof is based on two observations: the first is that the right-hand side of (3.6) does not depend on N and therefore it is easy to compute its limit as $N \to \infty$. The second observation is that the strong law of large numbers (SLLN) for Markov chains [100] can be applied to estimate the expectation on the right-hand side of (3.6). The detailed proof is in the Appendix C.1. □

An almost surely convergent estimator for $h(Y)$ can be easily derived using Theorem 3.1, simply by modeling the dependence of the current samples of Y on its own J'_{yy} past samples. This is equivalent to setting $K_{yx} = 0$. The difference between the two estimators, $\hat{h}(Y)$ and $\hat{h}(Y\|X)$, is the almost surely convergent estimator for DI from X to Y, $\hat{I}(X \to Y)$. This is stated in Theorem 3.2.

Theorem 3.2. *Let Assumptions 1 and 2 hold. The universal estimator for DI from time-series X to Y is*

$$\hat{I}(X \to Y) = \hat{h}(Y) - \hat{h}(Y\|X) \xrightarrow{a.s.} I(X \to Y). \tag{3.8}$$

Proof. We have from Theorem 3.1 $\hat{I}(X \to Y) = \hat{h}(Y) - \hat{h}(Y\|X) \xrightarrow{a.s.} h(Y) - h(Y\|X) = I(X \to Y)$. □

The DI estimator in Theorem 3.2 can be easily extended to estimate the causally conditioned directed information, $I(X \to Y\|W)$. First, $h(Y\|W)$ is estimated using Theorem 3.1. We now need to estimate $h(Y\|W, X)$. Let J_{yy}, K_{yw} and K_{yx} respectively denote the number of past samples of Y, W and X that influence the current sample of Y. Let us also assume the causal conditional likelihood $\mathrm{P}\left(y[n]|Y_{n-J_{yy}}^{n-1}, W_{n-K_{yw}+1}^{n}, X_{n-K_{yx}+1}^{n}\right)$ is known. A model-based and a data-driven approach to estimate this CCL is described in the subsequent section. Then Theo-

rem. 3.1 can be easily extended to show that

$$\hat{h}\left(Y\|W,X\right)=\frac{1}{N}\sum_{n=1}^{N}\left\{-\log\mathrm{P}\left(y[n]|Y_{n-J_{yy}}^{n-1},W_{n-K_{yw}+1}^{n},X_{n-K_{yx}+1}^{n}\right)\right\} \tag{3.9}$$

is an almost surely convergent estimate of $h\left(Y\|W,X\right)$. From (3.5), $\hat{I}\left(X\rightarrow Y\|W\right)$ is the difference between the estimates, $\hat{h}\left(Y\|W\right)$ and $\hat{h}\left(Y\|W,X\right)$. It is important to note that as the number of channels included in W increases, the computational complexity of the estimator also increases.

3.3.1 Statistical Significance Testing

The DI estimate, $\hat{I}\left(X\rightarrow Y\right)$, can be interpreted as the amount of causal information X contains about Y. It is, however, important to note that $\hat{I}\left(X\rightarrow Y\right)$ is estimated from N samples and is an estimate of the true value of DI from X to Y. The statistical significance of the causal connection from X to Y inferred from $\hat{I}\left(X\rightarrow Y\right)$ is calculated using an adaptation [97] of stationary bootstrap [98]. B stationary bootstrap samples of X, denoted by $X^{(b)}$, are generated using the algorithm described in [97] for $b=1,2,\cdots,B$. The DI from b^{th} stationary bootstrap sample $X^{(b)}$ to Y, denoted by $\hat{I}\left(X^{(b)}\rightarrow Y\right)$, is estimated using the proposed DI estimator. Note that there is no causal influence from any of these bootstrap samples to Y by construction. Therefore the B samples, $\hat{I}\left(X^{(b)}\rightarrow Y\right)$, for $b=1,2,\cdots,B$ are from the null hypothesis of no causal influence. The statistical significance is determined by the P-value [101]. P-value is the probability that DI estimate greater than or equal to $\hat{I}\left(X\rightarrow Y\right)$ can be observed under the null hypothesis of no causality from X to Y and is computed from the empirical distribution of $\hat{I}\left(X^{(b)}\rightarrow Y\right)$ for $b=1,2,\cdots,B$. If the P-value is less than a predetermined significance level δ, the null hypothesis of no causal con-

nection from X to Y is rejected. On the other hand, if the P-value is greater than δ, the null hypothesis cannot be rejected and the causal connection from X to Y is not statistically significant. Note that the empirical distribution of $\hat{I}\left(X^{(b)} \to Y\right)$ is concentrated around 0, since the DI between time-series that are not causally connected is zero. Therefore, when the actual DI estimate is large enough, the P-value will be less than δ and the statistical significance assessment is not required. However, statistical significance assessment is useful when the DI estimate is close to zero. The significance assessment described here is applied to the simulated examples in section 3.5 to identify the significant causal connections, particularly useful when the DI estimates are close to zero. The above discussion assumes CCL is known. The likelihood, however, must be estimated from data in practice. A model-based and a data-driven approach to estimate CCL is described in the following section.

3.4 Estimating Causal Conditional Likelihood

Estimating DI from X to Y using the proposed DI estimator in section 3.3 requires estimating two CCLs, $\mathrm{P}\left(y[n]|Y^{n-1}\right)$ and $\mathrm{P}\left(y[n]|Y^{n-1}, X^n\right)$, while estimating DI from Y to X causally conditioned on W requires estimating two CCLs, $\mathrm{P}\left(y[n]|Y^{n-1}, W^n\right)$ and $\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right)$. Let us focus on estimating $\mathrm{P}\left(y[n]|Y^{n-1}, X^n\right)$ for $n = 1, 2, \cdots, N$, which is required to estimate $\hat{h}\left(Y\|X\right)$. We will then describe how to extend this approach to estimate $\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right)$. The CCLs are estimated using either model-based or data-driven techniques. The choice between model-based and data-driven approaches is determined by the application from which data is recorded. For instance, the time-series signals obtained from electrophysiological recordings of brain or from stock markets are commonly modeled using MVAR models with Gaussian white noise. In this case, the CCL is easily estimated from the MVAR model

of the data. Usually the parameters of the model are unknown and several classical techniques to estimate the unknown parameters are described in [102]. On the other hand, using model-based approaches to estimate CCLs from data recorded from non-linear systems or systems without a prescribed linear model is non-trivial. This is because estimating the CCLs using model-based approach requires essentially inverting the nonlinear generative model, which is not trivial. Data-driven approaches do not have this limitation and are therefore preferred for nonlinear time-series data. A good review of the various data-driven algorithms that estimate probability distribution from data is provided in [5, 103]. The model-based and the data-driven CCL algorithm used in this thesis are described in the remainder of this section.

3.4.1 Model-based CCL Estimation

We will focus on estimating the CCL specifically for multivariate autoregressive process with Gaussian white noise in this thesis. Let the time-series X and Y be sampled from such processes. Then, the samples of Y can be expressed as

$$y[n] = \sum_{j=1}^{J_{yy}} \alpha_j y[n-j] + \sum_{k=1}^{K_{yx}} \beta_k x[n-k+1] + z[n], n = 1, 2, \cdots, N, \tag{3.10}$$

where z_n is the additive white Gaussian noise with zero mean and variance σ_z^2. Here α_j for $j = 1, 2, \cdots, J_{yy}$ and β_k for $k = 1, 2, \cdots, K_{yx}$ are the parameters of the model and J_{yy} and K_{yx} are the model orders representing how many past samples of Y and of X respectively influence the current sample of Y. It is easy to observe from (3.10) that

$$\mathrm{P}\left(y[n]|Y^{n-1}, X^n\right) \sim \mathcal{N}\Big(\sum_{j=1}^{J_{yy}} \alpha_j y[n-j] + \sum_{k=1}^{K_{yx}} \beta_k x[n-k+1], \sigma_z^2\Big). \tag{3.11}$$

The two model orders, J_{yy} and K_{yx}, and the parameter vector $\theta\left(J_{yy}, K_{yx}\right) = \left(\alpha_1, \cdots, \alpha_{J_{yy}}, \beta_1, \cdots, \beta_{K_{yx}}, \sigma_z^2\right)^{\mathrm{T}}$ are not known apriori and need to be estimated from the N observed samples of X and Y. The parameters and the model orders are estimated using a maximum likelihood (ML) estimator with minimum description length [104] penalty. ML estimator is known to be asymptotically consistent. Minimum description length is a model order selection procedure with good consistency properties [104] and proportional to $(J_{yy} + K_{yx})$. The optimal model orders $\left(\hat{J}_{yy}, \hat{K}_{yx}\right)$ are the solutions of the following problem:

$$\left(\hat{J}_{yy}, \hat{K}_{yx}\right) = \underset{(J_{yy}, K_{yx})}{\arg\min}\Big\{ -\tfrac{1}{N}\log \mathrm{P}\left(Y^N \| X^N; \hat{\theta}\left(J_{yy}, K_{yx}\right)\right) + \tfrac{J_{yy}+K_{yx}}{2N}\log N\Big\}, \quad (3.12)$$

where $\hat{\theta}\left(J_{yy}, K_{yx}\right)$ is the value of θ which minimizes the negative log-likelihood for a given $\left(J_{yy}, K_{yx}\right)$ and is obtained by solving

$$\hat{\theta}\left(J_{yy}, K_{yx}\right) = \underset{\theta}{\arg\min}\Big\{ -\tfrac{1}{N}\log \mathrm{P}\left(Y^N \| X^N; \theta\left(J_{yy}, K_{yx}\right)\right)\Big\}. \quad (3.13)$$

The ML estimation of θ for a given (J_{yy}, K_{yx}) in (3.13) is equivalent to the ML estimation of the parameters of a standard linear regression model [102], since the CCL is Gaussian distributed (3.11). The estimated parameters almost surely converge to the true parameter values [105] resulting in almost surely convergence of the proposed DI estimator. The desired CCL is obtained by substituting the solutions of (3.13), (3.12) in (3.11). The resultant CCL is then substituted in (3.7) to estimate $\hat{h}\left(Y \| X\right)$, which is further simplified to $\hat{h}\left(Y \| X\right) = \frac{1}{2}\log\left(2\pi e \hat{\sigma}_z^2\right)$, where $\hat{\sigma}_z^2$ is the estimate of the noise variance from (3.12), (3.13).

The MVAR model-based CCL estimation algorithm described above can be easily extended to estimate the CCLs required to estimate the causal conditional DI,

$\hat{I}\left(X \to Y \| W\right)$. Let us focus on estimating $\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right)$, which is required to estimate $\hat{h}\left(Y \| W, X\right)$. Assuming MVAR model, let J_{yy}, K_{yw}, K_{yx} respectively denote the number of past samples of Y,W, X that influence $y[n]$. Then for $n = 1, 2, \cdots, N$, the current sample of Y can be expressed as

$$y[n] = \sum_{j=1}^{J_{yy}} \alpha_j y[n-j] + \sum_{k=1}^{K_{yw}} \gamma_k w[n-k+1] + \sum_{l=1}^{K_{yx}} \beta_l x[n-l+1] + z[n]. \tag{3.14}$$

The only difference with (3.10) are the extra terms of the time-series W. As a result, the CCL will still be Gaussian distributed with same variance as the distribution in (3.11) and whose mean contains the extra terms corresponding to the samples of W. The unknown parameters under this model are $\alpha_j, \gamma_k, \beta_l$ for $j = 1, \cdots, J_{yy}$, $k = 1, \cdots, K_{yw}$,$l = 1, \cdots, K_{yx}$ and the model orders J_{yy}, K_{yw}, K_{yx}. Maximum likelihood with minimum description length penalty can be used to estimate these parameters similarly. The resulting parameter estimates can then be used to calculate the CCL, which is substituted in (3.9) to estimate $\hat{h}\left(Y \| W, X\right)$.

3.4.2 Data-driven CCL Estimation

Let J_{yy} and K_{yx} denote the number of past samples of Y and X that influence the current sample of Y. Then the CCL $\mathrm{P}\left(y[n]|Y^{n-1}, X^n\right)$ is same as $\mathrm{P}\left(y[n]|Y_{n-J_{yy}}^{n-1}, X_{n-K_{yx}+1}^{n}\right)$ and can be written as

$$\mathrm{P}\left(y[n]|Y_{n-J_{yy}}^{n-1}, X_{n-K_{yx}+1}^{n}\right) = \frac{\mathrm{P}\left(Y_{n-J_{yy}}^{n}, X_{n-K_{yx}+1}^{n}\right)}{\mathrm{P}\left(Y_{n-J_{yy}}^{n-1}, X_{n-K_{yx}+1}^{n}\right)}. \tag{3.15}$$

The joint distribution $\mathrm{P}\left(Y_{n-J_{yy}}^{n}, X_{n-K_{yx}+1}^{n}\right)$ of J_{yy}+1 and K_{yx} consecutive samples of Y and X respectively is learned using kernel density estimator [103] with Gaussian kernels. This estimator is implemented in the 'ks' package in R [75]. The true

values of (J_{yy}, K_{yx}) are not known and should be estimated. The joint density is learned for different values of J_{yy} and K_{yx} and the optimal values $\left(\hat{J}_{yy}, \hat{K}_{yx}\right)$ are those that maximize the likelihood. The desired CCL is then estimated by substituting $\mathrm{P}\left(Y^n_{n-\hat{J}_{yy}}, X^n_{n-\hat{K}_{yx}+1}\right)$ in (3.15). The denominator in (3.15) marginalizes the joint distribution in numerator of (3.15) over $y[n]$. This marginalization is implemented by approximating the integral with a Riemann sum of the distribution over a partition of the range of $y[n]$. Note that the convergence of the estimated CCL to the true CCL depends on the underlying true data distribution [106]. $\hat{h}\left(Y\|X\right)$ is obtained by substituting the estimated CCL in (3.7).

The data-driven CCL estimation algorithm described above can be extended to estimate $\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right)$ as well. Let J_{yy}, K_{yw}, K_{yx} respectively denote the number of past samples of Y, W, X that influence $y[n]$. Then

$$\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right) = \frac{\mathrm{P}\left(Y^n_{n-J_{yy}}, W^n_{n-K_{yw}+1}, X^n_{n-K_{yx}+1}\right)}{\mathrm{P}\left(Y^{n-1}_{n-J_{yy}}, W^n_{n-K_{yw}+1}, X^n_{n-K_{yx}+1}\right)}. \tag{3.16}$$

The joint distribution in the numerator can be similarly estimated using kernel density estimator [103] with Gaussian kernels using 'ks' package [75]. Note that the optimal values of the model-orders J_{yy}, K_{yw}, K_{yx} are those that maximize the likelihood. The denominator in (3.16) is then obtained by marginalizing the distribution in the numerator similarly. The resultant numerator and denominator probabilities are substituted in (3.16) to estimate $\mathrm{P}\left(y[n]|Y^{n-1}, W^n, X^n\right)$, which is further substituted in (3.9) to estimate $\hat{h}\left(Y\|X, W\right)$.

The model-based and data-driven CCL algorithms described above can be easily modified to estimate $\mathrm{P}\left(y[n]|Y^{n-1}\right)$, which is required to estimate $\hat{h}\left(Y\right)$. $\mathrm{P}\left(x[n]|Y^{n-1}\right)$ is obtained from either model-based or data-driven CCL by modeling the depen-

dence of the current sample of Y just on its own past samples. $I(X \to Y)$ and $I(X \to Y \| W)$ can now be estimated using the estimator proposed in section 3.3.

The DI estimator obtained by using the proposed estimator in Theorem. 3.2 with model-based CCL and data-driven CCL estimation algorithms will henceforth be referred to as model-based and data-driven DI estimator respectively. If data is assumed to be drawn from MVAR model with Gaussian white noise, then model-based DI will be referred to as MVAR model-based DI estimator. Note that model-based approach is not restricted to just MVAR models, it is feasible for all those models from which we can estimate the appropriate causal conditional likelihoods parametrically. We focused on MVAR with Gaussian white noise in this thesis because ECoG is commonly modeled using this model in connectivity studies [50,52]. The performance of both the proposed DI estimators on simulated time-series data is demonstrated in the following section.

3.5 Performance on Simulated Data

In this section, the performance of the proposed DI estimators is demonstrated using simulated data generated from six models - two node bidirectional linear (section 3.5.1) and nonlinear (section 3.5.2) causal network whose true connectivity is depicted in Fig. 3.1a, a two node unidirectional noisy chaotic polynomial (section 3.5.3) causal network whose true connectivity is shown in Fig. 3.1b, four node linear (section 3.5.4) and nonlinear (section 3.5.5) causal network whose true connectivity is depicted in Fig. 3.1c and a six node linear (section 3.5.6) causal network depicted in Fig. 3.1d. A directed arrow in Fig. 3.1 represents a causal connection. The causal connection between two nodes, say from node A to B in Fig. 3.1c, implies $I(A \to B) > 0$ or equivalently, that the past samples of A have some information about the current

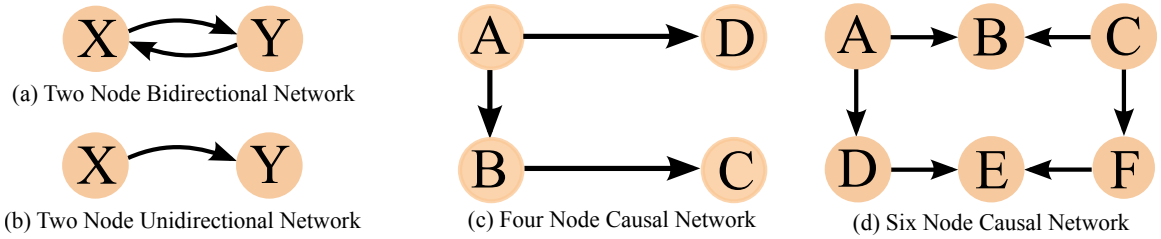


Figure 3.1 : The true causal connectivity graphs of the simulated data models used to validate the performance of the proposed model-based and data-driven DI estimators.

sample of B. We also compared the performance of the proposed DI estimators with the standard Granger causality (GC) [79]. GC estimate is obtained from MVGC toolbox [96]. Let us now describe the performance of the proposed DI estimators on the six models considered in detail.

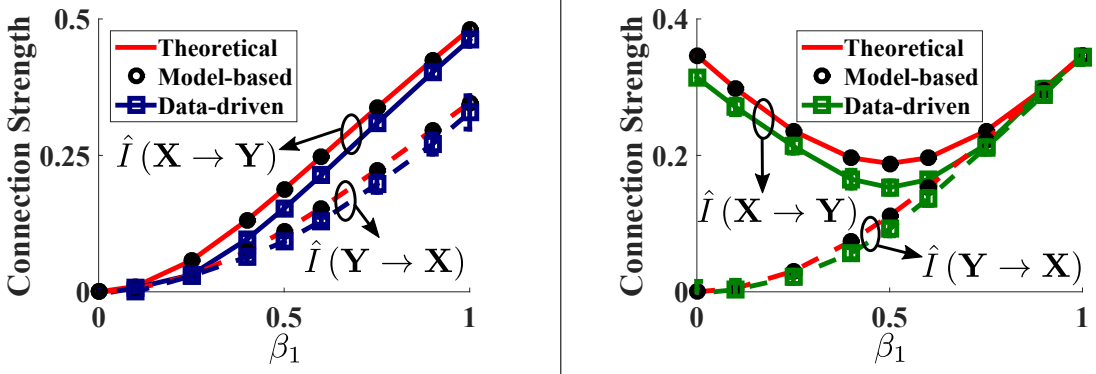
3.5.1 Two Node Bidirectional Linear Causal Network

Consider two time-series X and Y causally connected as shown in Fig. 3.1a. The time-series Y is generated from

$$y[n] = \beta_1 x[n] + \beta_2 x[n-1] + z[n], \text{ for } n = 1, 2, \cdots, N, \tag{3.17}$$

where $x[n]$ and $z[n]$ are sampled from an i.i.d Gaussian distribution with zero mean and variance σ_x^2, σ_z^2 respectively. The samples of X and Z are independent. The true value of the DI between X and Y in both directions is used to benchmark the performance of the proposed model-based and data-driven DI estimators.

Let us first look at the true value of DI for the model by (3.17) in two special cases. When $\beta_1 = 1, \beta_2 = 0$, (3.17) reduces to $y[n] = x[n] + z[n]$, and it is obvious that both X and Y have equal causal information about each other. It is easy to see that $I(X \to Y) = I(Y \to X) = I(X;Y) = C$, where $I(X;Y)$ is the mutual



(a) $\beta_2 = \beta_1$ (b) $\beta_2 = 1 - \beta_1$

Figure 3.2 : DI estimates and their standard deviation for the two node network (in Fig. 3.1a) generated from a linear model (3.17) using analytical expression (3.18), proposed model-based and data-driven DI estimators for different values of causal strength quantified by (β_1, β_2). The DI estimates are plotted against β_1 with $\beta_2 = \beta_1$ in Fig. 3.2a and with $\beta_2 = 1 - \beta_1$ in Fig. 3.2b.

information between X and Y and $C = \frac{1}{2}\log\left(1 + \frac{\sigma_x^2}{\sigma_z^2}\right)$. The other special case occurs when $\beta_1 = 0, \beta_2 = 1$ and in this case (3.17) reduces to $y[n] = x[n-1] + z[n]$. In this case, X has causal information about Y, while Y has no causal information about X. More precisely, $I(X \to Y) = I(X;Y) = C$ and $I(Y \to X) = 0$. For the remaining case of non-zero β_1, β_2, the analytical expressions for DI are

$$I\big(X \to Y\big)x[n-2] = \frac{1}{2}\log\left(\frac{|\beta_1\beta_2|\sigma_x^2}{\sigma_z^2}\right) + \frac{1}{2}\cosh^{-1}\left(\frac{\left(\beta_1^2+\beta_2^2\right)\sigma_x^2+\sigma_z^2}{2|\beta_1\beta_2|\sigma_x^2}\right),$$
$$I\big(Y \to X\big) = \frac{1}{2}\log\left(1 + \frac{\beta_1^2\sigma_x^2}{\sigma_z^2}\right). \tag{3.18}$$

The derivation of (3.18) uses the tridiagonal matrix determinant from [107] and is given in Appendix C.2. Note from (3.18) that DI from Y to X does not depend on β_2. It is because the uncertainty in the current sample of X does not depend on β_2, when causally conditioned on the past of X and Y.

The DI from X to Y and vice versa is estimated from $N = 10^5$ samples of X

and Y generated with $\sigma_x^2 = 1$, $\sigma_z^2 = 1$ using the proposed model-based and data-driven DI estimators. The model-based DI estimator assumes that the time-series are modeled by a MVAR model with Gaussian white noise, whereas the data-driven CCL estimator does not impose any model assumptions on the data. Assuming X, Y are from a MVAR process and when $x[n]$ is included in the past samples of X, Granger causality estimate from X to Y is equal to twice the MVAR model-based DI estimate from X to Y and vice versa [94]. We therefore do not show the GC estimates for linear MVAR models with Gaussian white noise. GC estimates are plotted only for nonlinear simulated models in this chapter.

Fig. 3.2 plots directed information values obtained from the analytical expression in (3.18), $\hat{I}\left(X \to Y\right)$ and $\hat{I}\left(Y \to X\right)$ from the proposed model-based and data-driven DI estimators for different values of $\beta_1 \in (0, 1)$. The corresponding curves are respectively referred to as theoretical, model-based and data-driven. For the model-based and data-driven curves in Fig. 3.2, multiple datasets of X, Y are generated using different seeds for the random number generator. The mean and the standard deviation of the resultant estimates are plotted in Fig. 3.2. The average standard deviation across all (β_1, β_2) in Fig. 3.2 is about 0.003 and 0.01 for the model-based and data-driven DI estimators respectively. $\beta_2 = \beta_1$ in Fig. 3.2a and $\beta_2 = 1 - \beta_1$ in Fig. 3.2b.

When $\beta_1 = \beta_2$, a larger β_1 implies a stronger causal connection between X and Y and this should result in a larger DI. This expected trend is observed in Fig. 3.2a. This implies that DI tracks the strength of the causal connection. Also in the corner case of $\beta_1 = \beta_2 = 0$, DI is zero in both directions as expected. In Fig. 3.2b, DI estimates in the corner cases of $\beta_1 = 0$, $\beta_2 = 1$ and $\beta_1 = 1$, $\beta_2 = 0$ match with the analytical expression as expected. Also as β_1 increases from 0 to 1, the causal information Y has about X increases, and DI tracks this. This is demonstrated by

observing that $\hat{I}(Y \to X)$ increases with β_1 in Fig. 3.2b. Finally, it is clear from Fig. 3.2 that the model-based estimate matches the correct value of DI estimate from (3.18) and the data-driven estimator follows the true value of DI. This validates the accuracy of the proposed DI estimators. For this MVAR model with Gaussian white noise, the model-based DI estimator clearly performs better than the data-driven DI estimator and also has a lower run-time. We therefore use the MVAR model-based estimator to estimate DI between data modeled by MVAR processes with Gaussian white noise, instead of using the data-driven estimator.

The adaptation of stationary bootstrap algorithm described earlier is used to assess the significance of the inferred causal connections for different values of (β_1, β_2). We observed that the null hypothesis of no causality from Y to X cannot be rejected for $\beta_1 \in \{0, 0.1\}$ (P-value $> \delta = 0.05$) and can be rejected at all other points (P-value $< \delta$) in Fig. 3.2. This is not surprising since $\hat{I}(Y \to X)$ is small for $\beta_1 \in \{0, 0.1\}$ and hence did not result in a significant causal connection from Y to X. Similarly, we observed that statistically significant causal connection from X to Y does not exist for $\beta_1 = 0, \beta_2 = 0$ (P-value $> \delta$) and exits at all other points (P-value $< \delta$) in Fig. 3.2. This once again confirms our intuition that only large positive values of DI imply a statistically significant causal connection.

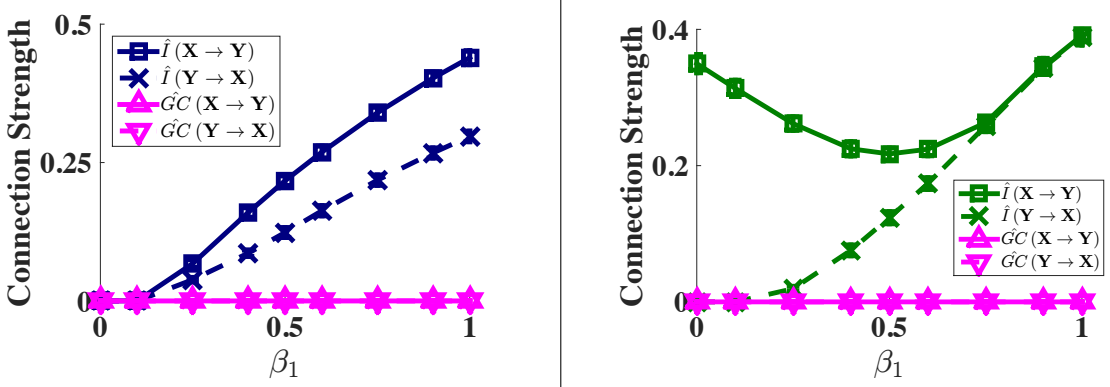
3.5.2 Two Node Bidirectional Nonlinear Causal Network

Now, consider time-series X and Y causally connected as shown in Fig. 3.1a and are generated according to

$$y[n] = \beta_1 x[n]^2 + \beta_2 x[n-1]^2 + z[n], \text{ for } n = 1, 2, \cdots, N, \tag{3.19}$$

where $x[n]$ and $x[n]$ are sampled from an i.i.d Gaussian distribution with zero mean and variance σ_x^2, σ_z^2 respectively. Also, the samples of X and Z are independent. It is very non-trivial to estimate $\hat{I}(X \to Y)$ and $\hat{I}(Y \to X)$ using model-based DI estimator. This is because estimating $p\left(x[n]|X_1^{n-1}, Y_1^n\right)$ and $p\left(y[n]|Y_1^{n-1}\right)$ requires essentially inverting the non-linear, non-Gaussian generative model in (3.19) and this is very hard even for this simple nonlinear model. These two probability densities are required to estimate $\hat{h}(X\|Y)$ and $\hat{h}(Y)$ respectively. Therefore we only use the proposed data-driven DI estimator to estimate the DI from X to Y and vice versa. However, we can always assume that the data from the model in (3.19) comes from a MVAR model with Gaussian noise, which is incorrect and estimate DI using the proposed MVAR model-based DI estimator. The resulting DI estimate will be half of the Granger causality estimate between these two time-series, $\hat{GC}(X \to Y)$ and $\hat{GC}(Y \to X)$. Note that GC also assumes the data is generated from a MVAR process even though it is incorrect. We will now compare the performance of data-driven DI and GC estimates on this model.

Directed information and Granger causality between X and Y in both directions is estimated from $N = 10^5$ samples generated with $\sigma_x^2 = 1$, $\sigma_z^2 = 1$ for different values of (β_1, β_2) and plotted in Fig. 3.3. The DI and GC estimates are plotted for $\beta_2 = \beta_1$ and $\beta_2 = 1 - \beta_1$ in Fig. 3.3a and Fig. 3.3b respectively. For each (β_1, β_2), multiple datasets of X, Y are generated with different random number generator seeds. The mean and the standard deviation of the resultant data-driven DI and GC estimates are plotted in Fig. 3.3. The average standard deviation across all (β_1, β_2) of the data-driven DI and GC estimates is 0.01 and 1.8×10^{-5} respectively. In addition, the search space of the model order used by the Granger causality estimator is up to 20, i.e, $J_{yy}, K_{yx} \in [1, 20]$. In Fig. 3.3a, $\hat{I}(X \to Y)$ increases with β_1 as expected. DI estimates



(a) $\beta_2 = \beta_1$ (b) $\beta_2 = 1 - \beta_1$

Figure 3.3 : Data-driven DI and GC estimates, along with standard deviation of the estimates, for the two node network (depicted in Fig. 3.1a) generated from the nonlinear model (3.19) for different values of causal strength quantified by (β_1, β_2). The estimates are plotted against β_1 with $\beta_2 = \beta_1$ in Fig. 3.3a and with $\beta_2 = 1 - \beta_1$ in Fig. 3.3b.

also behave as expected in the corner cases of $(\beta_1, \beta_2) = (0, 1)$ and $(1, 0)$ in Fig. 3.3b. $\hat{I}(Y \to X)$ increases with β_1 as expected. This once again demonstrates that DI tracks the strength of causal connections. On the other hand, Granger causality estimates in both directions are almost zero (of the order of 10^{-5}), indicating that Granger causality cannot detect the causal connections in nonlinear models.

The statistical significance of the inferred causal connections by DI and GC estimates for different values of (β_1, β_2) in Fig. 3.3 is assessed using the stationary bootstrap algorithm described in section 3.3. Using DI, the null hypothesis of no causality from Y to X cannot be rejected for $(\beta_1, \beta_2) \in \{(0, 0), (0, 1)(0.1, 0.1), (0.1, 0.9)\}$ and from X to Y cannot be rejected for $(\beta_1, \beta_2) = (0, 0)$ (P-value $> \delta = 0.05$) in Fig. 3.3. At all other points in Fig. 3.3, the null hypothesis of no causality can be rejected (P-value $< \delta$) using DI estimates. This once again confirms our intuition that large values of DI imply a statistically significant causal connection. For GC, the null hypothesis of no causality cannot be rejected at all points in Fig. 3.3 implying that

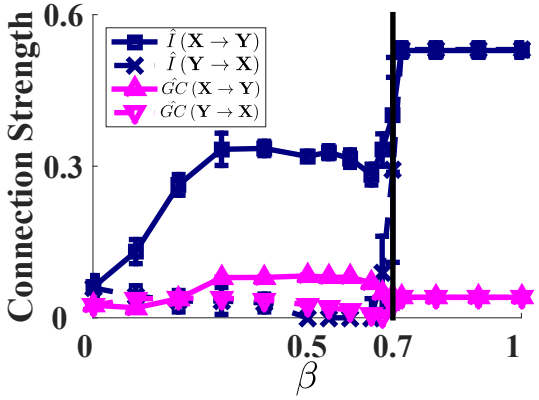


Figure 3.4 : Data-driven DI estimates and GC estimates, along with standard deviation of the estimates, for two node unidirectional network in Fig. 3.1b generated from noisy chaotic polynomial map (3.20) for different values of the coupling parameter β.

GC could not find statistically significant causal connections in nonlinear models. This example proves that DI is a more general causal connectivity metric that is not restricted to some particular models.

3.5.3 Two Node Unidirectional Noisy Chaotic Polynomial Map

We now consider two unidirectionally coupled time-series X and Y whose underlying causal connectivity is shown in Fig. 3.1b. The time-series X and Y are generated from a noisy chaotic polynomial map [108] according to

$$\begin{aligned} x[n] &= 1.4 - x[n-1]^2 + 0.3y[n-2], \\ y[n] &= 1.4 - (\beta x[n-1] + (1-\beta)\, y_{n-1})\, y_{n-1} + 0.3y[n-2], \end{aligned} \tag{3.20}$$

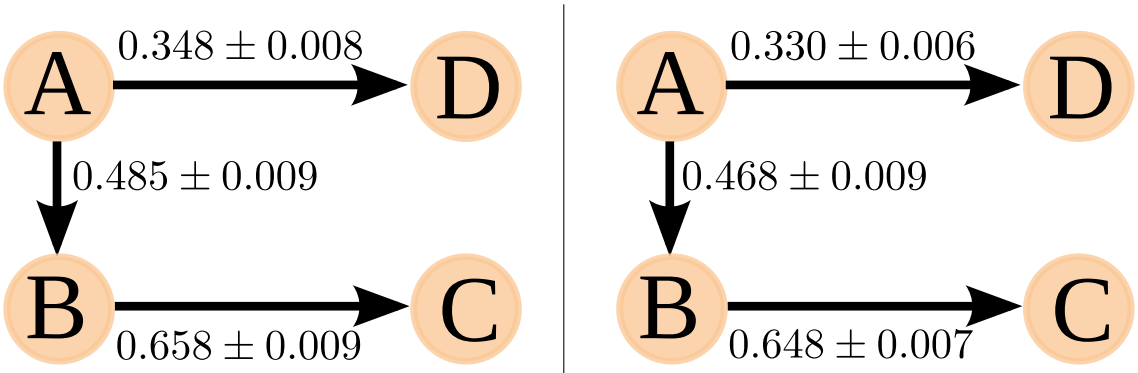
where β controls the amount of causal information flowing from X to Y. The initial two samples, $x[1], x[2], y[1], y[2]$ are randomly chosen. The two time-series become completely synchronized for $\beta > 0.7$. Gaussian i.i.d measurement noise of variance

0.01 is added to both time-series X and Y. For $\beta \in [0, 0, 7)$, strength of the causal connection from X to Y should increase with β and there is no causal connection from Y to X. For $\beta \in (0.7, 1]$, since both time-series are completely synchronized and because of the measurement noise, there is a non-zero equally strong causal connection in both directions. In the absence of measurement noise for $\beta \in (0.7, 1)$, $x[n] = z[n]$ leading to causal conditional entropy estimate of negative infinity and a DI estimate of infinity. The intuition behind this is that once the past of X is known, there is no uncertainty left in Y. On the other hand, GC estimates in the synchronized range will be close to zero because the past of X used by GC (unlike DI, GC does not include $y[n]$ in the past of X) does not contain any predictive information about $y[n]$ resulting in a GC estimate of zero from X to Y. Note that it is very non-trivial to apply model-based DI on this model because of the same reasons outlined in the previous simulated nonlinear model. We therefore only compare the performance of data-driven DI and GC estimates on this model.

DI and GC in both directions is estimated from $N = 10^5$ samples of X and Y (after discarding the initial transient points) for different values of $\beta \in [0, 1]$ and plotted in Fig. 3.4. For each β, the time-series are generated from (3.20) using different seeds of the random number generator. The mean and the standard deviation of the resulting data-driven DI and GC estimates are plotted in Fig. 3.4. The average standard deviation across all β for the data-driven DI and GC estimates is 0.03 and 0.001 respectively. The standard deviation was largest at $\beta = 0.7$, implying that it is very hard to estimate at the boundary before and after complete synchronization. In addition, the search space of the model order used by the Granger causality estimator is up to 20, i.e, $J_{yy}, K_{yx} \in [1, 20]$. The DI estimate is obtained by subtracting two non-negative numbers and it can sometimes be a small negative number because of the

inaccuracies in estimation algorithms or insufficient data or violation of stationarity assumptions [88] and in those cases, we reset the DI estimate to be zero. For instance, the largest negative DI estimate we obtained for this model is -0.06 from Y to X at $\beta = 0.6$ and we reset this estimate to 0. It is clear from Fig. 3.4 that DI estimates behave as expected. DI from X to Y increase as β goes from 0 to 1. On the other hand, the DI estimates from Y to X are very small numbers for $\beta < 0.7$ and then there is a sudden jump in this estimate after $\beta > 0.7$. This jump is because the time-series get synchronized for $\beta > 0.7$. On the other hand, GC estimates in both directions are small positive numbers (when compared to DI estimates) for the whole range and become equal in value in the synchronized range of $\beta > 0.7$.

The statistical significance of the causal connections inferred by DI and GC estimates is assessed using the adaption of stationary bootstrap. The null hypothesis of no causality using DI estimates from Y to X cannot be rejected for $\beta < 0.7$ and cannot be rejected for the connection from X to Y for $\beta < 0.1$. This implies DI correctly identifies the presence of causal connection from X to Y for all $\beta \geq 0.1$ and the absence of causal connection from Y to X for $\beta < 0.7$. It can also differentiate causally independent time-series ($\beta = 0$) and completely identical time series ($\beta \in (0.7, 1]$). On the other hand, the null hypothesis of no causality cannot be rejected only for $\beta = 0$ using GC estimates. This implies GC identifies the presence of a causal connection in both directions for all non-zero β, which is incorrect. This example also shows DI correctly infers causal connectivity from nonlinear models.



(a) Model-based Estimator (b) Data-driven Estimator

Figure 3.5 : The causal network along with connection strengths between the four MVAR processes simulated from (3.21) estimated by the MVAR model-based DI and the data-driven DI estimators. The true causal connectivity graph between these four time-series is depicted in Fig. 3.1c. It is clear that both DI estimators correctly infer the underlying causal network.

3.5.4 Four Node Linear Causal Network

Now, consider the four node causal network depicted in Fig. 3.1c. The four time series A, B, C and D are generated according to

$$\begin{aligned} b[n] &= a[n-1] + a[n-2] + z^b[n], \quad c_n = b[n-1] + z^c[n], \\ d[n] &= a[n-2] + z^d[n], \quad \text{for} \quad n = 1, 2, \cdots, N, \end{aligned} \tag{3.21}$$

where $a[n]$, $z^b[n]$, $z^c[n]$ and $z^d[n]$ are sampled from an i.i.d Gaussian distribution with zero mean and unit variance. In this network, A influences C indirectly through B. This is an example of an 'indirect' causal connection, in contrast with the connection from A to B, which is a 'direct' causal connection. DI estimate between pairs of time-series cannot differentiate between 'direct' and 'indirect' causal connections [82]. For instance, the DI estimate from A to C is positive, even though A does not directly influence C, but causally influences C via B. A thorough discussion on the direct

and indirect influences for point processes is in [82] and is directly applicable here. Following the approach taken in [94, 109], the 'direct' causal influence from A to C is non-zero, if and only if $I\left(A \to C \| B, D, \right) > 0$. However, estimating the causally conditioned DI when the number of channels recorded from is large (of the order of hundred's) is difficult because of the curse of dimensionality [5]. To overcome this, the pairwise DI is first estimated between all pairs of channels. The indirect influences are then resolved by first estimating only the required causal DI between two processes, conditioned on one more process. Then if required, the causal DI between two processes, conditioned on two more processes, is estimated and so on. The termination condition is determined by the desired degree of 'directness' in the inferred causal network. In this simulated example, we are interested in recovering the true 'direct' causal network depicted in Fig. 3.1c.

To infer the true causal network, DI is estimated between these four time-series using both MVAR model-based and data-driven DI estimators. Model-based DI estimator assumes the data is generated from a linear causal MVAR model, whereas data-driven DI estimator does not impose any parametric model assumptions on the data. The data is generated from (3.21) using 20 different seeds to generate the Gaussian noise and the resultant estimates are averaged. We will first describe the performance using model-based DI estimator.

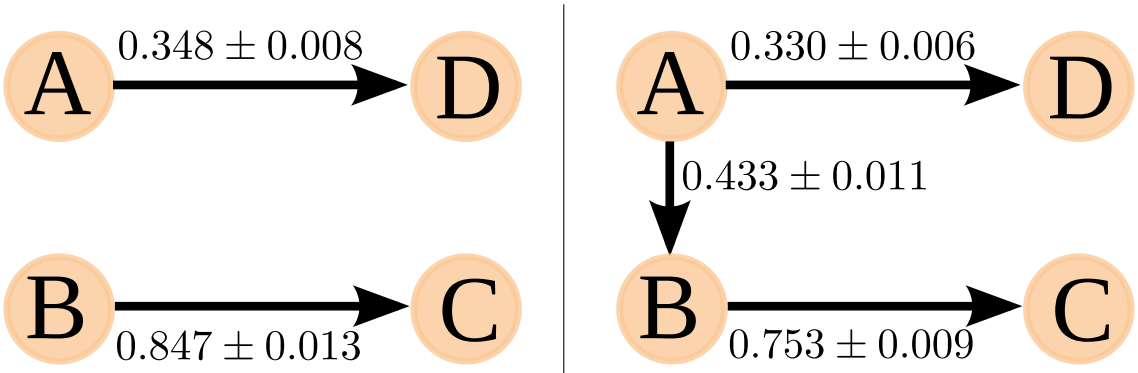
Model-based DI estimator is used to estimate the pairwise DI between all pairs of these four nodes, resulting a 4×4 matrix with zeros on the diagonal. We found that $\hat{I}\left(A \to B\right) = 0.485 \pm 0.009, \hat{I}\left(A \to C\right) = 0.314 \pm 0.009 \text{and} \hat{I}\left(B \to C\right) = 0.658 \pm 0.009$. To determine if there is an indirect causal connection from A to C or from B to C, we estimated $\hat{I}\left(A \to C \| B\right)$ and $\hat{I}\left(B \to C \| A\right)$ using the model-based causally conditioned DI estimator described in section 3.3, 3.4.1. We found that $\hat{I}\left(A \to C \| B\right) = 0$

and $\hat{I}\left(B \to C \| A\right) = 0.344 \pm 0.009$. Therefore, A to C is an 'indirect' connection via B. Causally conditional DIs are estimated till the network is completely resolved and free of any indirect influences. The estimated causal network along with the strength and the standard deviation of the estimated causal connections is depicted in Fig. 3.5a. It is clear from Fig. 3.5a and Fig. 3.1c that model-based DI estimator infers the true causal network correctly.

We now use the data-driven DI estimator to infer the true causal network. The pairwise DI is estimated between all pairs of these four nodes using the data-driven estimator, resulting in a 4×4 matrix with zeros on the diagonal. Using this DI estimator, we find that $\hat{I}\left(A \to B\right) = 0.468 \pm 0.009$, $\hat{I}\left(A \to C\right) = 0.296 \pm 0.004$ and $\hat{I}\left(B \to C\right) = 0.648 \pm 0.008$. To identify the presence of any indirect connections, we estimated $\hat{I}\left(A \to C \| B\right)$ and $\hat{I}\left(B \to C \| A\right)$ using the model-based causally conditioned DI estimator described in section 3.3, 3.4.2. We found that $\hat{I}\left(A \to C \| B\right) = 0$ and $\hat{I}\left(B \to C \| A\right) = 0.273 \pm 0.009$. Therefore, A to C is an 'indirect' connection via B. This procedure is continued to identify and remove all indirect causal connections. The resultant estimated direct causal network is depicted in Fig. 3.5b. It is clear that data-driven DI also recovers the true network correctly. Moreover, it is clear from Fig. 3.5 that for this model, both model-based and data-driven DI estimators correctly infer the underlying causal network, which is not surprising since the underlying model is a linear MVAR model.

3.5.5 Four Node Nonlinear Causal Network

We now use a nonlinear model to generate the four time-series A, B, C and D whose underlying causal connectivity graph is depicted in Fig. 3.1c. N samples from the



(a) Model-based Estimator (b) Data-driven Estimator

Figure 3.6 : The causal network along with connection strengths between the four time-series simulated from (3.22) estimated by the MVAR model-based DI and the data-driven DI estimators. The true causal connectivity graph between these four time-series is depicted in Fig. 3.1c. It is clear that unlike MVAR model-based estimator, the data-driven estimator correctly infers the underlying causal connectivity graph.

four time-series are generated according to

$$b[n] = a[n-1]^2 + a[n-2]^2 + z^b[n], \quad c[n] = b[n-1] + z^c[n],$$
$$d[n] = a[n-2] + z^d[n], \quad \text{for} \quad n = 1, 2, \cdots, N, \tag{3.22}$$

where $a[n]$, $z^b[n]$, $z^c[n]$ and $z^d[n]$ are sampled from an i.i.d Gaussian distribution with zero mean and unit variance. The only difference with the model in section 3.5.4 is that the causal connection from A to B is now nonlinear.

First, we infer the true causal connectivity for this model using the MVAR model-based DI estimator. This DI estimator assumes that the data is drawn from a linear MVAR model, which is not true for this model. It is clear from (3.22) that the time-series B is not generated from a linear MVAR model. Pairwise DI is estimated using this model between all pairs of these four time-series resulting in a 4×4 matrix with zeros on the diagonal. The only significant causal connections estimated by

the model-based DI estimator are from B to C and from A to D. This process is repeated for data generated using 20 different seeds and the resultant DI estimates are averaged. We find that $\hat{I}\left(B \to C\right) = 0.847 \pm 0.013$ and $\hat{I}\left(A \to D\right) = 0.348 \pm 0.008$. It is also clear that there are no indirect connections to resolve in this case. The underlying causal connectivity graph estimated by the model-based DI estimator is depicted in Fig. 3.6a. It is clear from this figure that model-based DI estimator could not recover this true network correctly. This is not surprising since the MVAR model-based estimator can only identify linear causal connections and cannot identify the nonlinear causal connections. As result, the connection from A to B is not identified by the model-based DI estimator.

We now use data-driven DI estimator to infer the causal connectivity from the simulated data. The pairwise DI is estimated between all pairs of these four nodes using the data-driven estimator, resulting in a 4×4 matrix with zeros on the diagonal. In contrast to the model-based DI estimator, we find that DI from A to B estimated using data-driven DI is nonzero. Specifically, we find that $\hat{I}\left(A \to B\right) = 0.433 \pm 0.011$. In addition, we also find that $\hat{I}\left(A \to C\right) = 0.320 \pm 0.010$ and $\hat{I}\left(B \to C\right) = 0.753 \pm 0.009$. To eliminate indirect causal connections, we estimated $\hat{I}\left(A \to C \| B\right) = 0$ and $\hat{I}\left(B \to C \| A\right) = 0.262 \pm 0.037$. Therefore, A to C is an 'indirect' connection via B. This procedure is continued to identify and remove all indirect causal connections. The resultant estimated direct causal network is depicted in Fig. 3.6b. It is clear that data-driven DI estimator recovers the true network correctly, while the model-based DI estimator could not infer the true causal network correctly.

3.5.6 Multinode Linear MVAR Causal Network

Now, consider the final model of the six node causal network depicted in Fig. 3.1d. We generate the six time-series, **A**, **B**, **C**, **D**, **E** and **F**, from a MVAR process (3.10). This network has indirect connections, for instance, **A** influences **E** indirectly through **D**. In this simulated example, we are interested in recovering the true 'direct' causal network depicted in Fig. 3.1b.

To infer the true causal network, $N = 10^5$ samples from the six time-series, **A**, **B**, **C**, **D**, **E** and **F** are generated from a MVAR process. Each process is dependent on its own past. The model order for the causal influence of one process on another is chosen from the set $\{0, 5, 10, 15\}$, where a model order 0 means no causal influence. The first filter coefficient of the causal connection between any two processes is set to zero, i.e., $\beta_1 = 0$. The other filter coefficients are generated from a uniform random number generator. The standard deviation of the white Gaussian noise used to generate **A**, **C** is 1 and to generate the remaining time-series is 0.1. For instance, the current sample of **D** is generated according to (3.10) and is influenced by 15 past samples of **D** and 5 past samples of **A**. The parameters of (3.10) used to generate samples of **D** are

$$\left\{\alpha_1^D, \cdots, \alpha_{10}^D\right\} = \{.01, .04, .09, .08, .10, .07, .00, .08, .09, .07, .08, .07, .04, .07, .02\},$$
$$\left\{\beta_1^{AD}, \cdots, \beta_5^{AD}\right\} = \{0, .28, .68, .66, .16\}, \sigma_z^D = 0.1.$$

The pairwise DI between every pair of processes is estimated from the $N = 10^5$ samples using the proposed MVAR model-based DI estimator. The causal connections with very small DI estimates are not statistically significant. Then the causally conditioned DI is estimated whenever necessary, to remove 'indirect' influences. For example, $\hat{I}(\mathbf{B} \to \mathbf{D}) = 0.24$, $\hat{I}(\mathbf{A} \to \mathbf{B}) = 0.48$ and $\hat{I}(\mathbf{A} \to \mathbf{D}) = 1.64$. To check

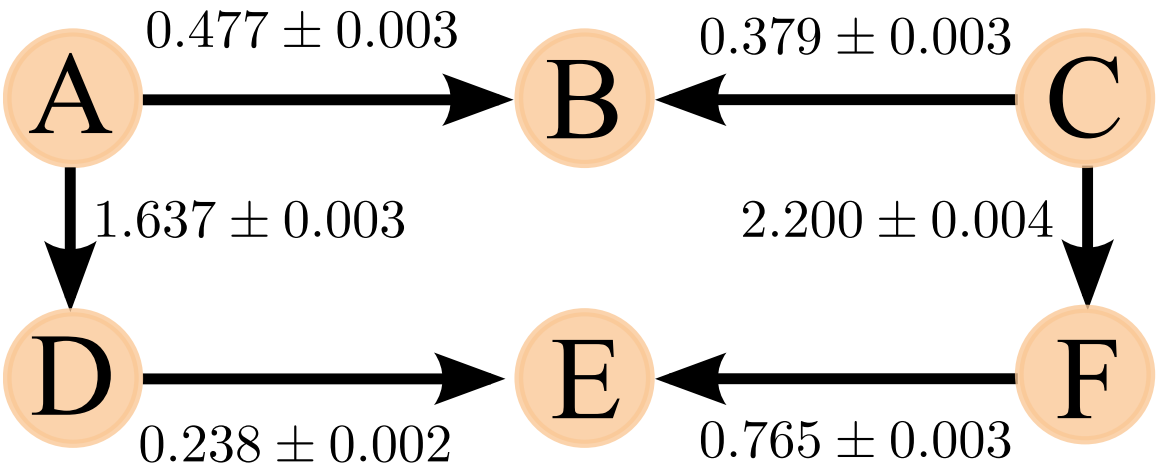


Figure 3.7 : Estimated causal network along with connection strengths between six simulated MVAR processes using DI, which matches with the true network in Fig. 3.1c.

whether **B** to **D** is an indirect connection via **A**, we estimated $\hat{I}(\mathbf{B} \to \mathbf{D} \| \mathbf{A})$ and found it to be 0. Therefore, **B** to **D** is an 'indirect' connection via **A**. Conditional DIs are estimated till the network is completely resolved and free of any 'indirect' influences. This procedure is repeated 20 times, each time using a different seed to generate the Gaussian noise and the resultant estimates are averaged. The estimated 'direct' causal network, along with the strength and the standard deviation of the estimated causal connections, is depicted in Fig. 3.7. It is clear from Fig. 3.1b and Fig. 3.7 that DI infers the true causal network correctly.

The six diverse simulated models considered in this section demonstrate that the DI correctly infers the presence and tracks the strength of a causal connection - large values of DI imply a strong causal connection and vice versa. Using stationary bootstrap, we also showed that only large positive DI estimates correspond to statistically significant causal connections. We also observed that model-based DI estimator cannot identify nonlinear causal connections, whereas data-driven DI estimator can correctly identify both linear and nonlinear causal connections. We only consider the

large DI estimates (large compared to the rest of the causal connectivity graph) since they only imply a significant causal connection.

3.6 Discussion and Conclusions

An almost surely convergent MVAR model-based and data-driven estimators for DI are introduced in this chapter. Linear causal interactions between two time-series can be quantified using MVAR model-based DI estimator, whereas both linear and nonlinear causal interactions are quantified by data-driven DI estimator. The resultant DI estimates can be used to infer whether the data has (1) linear causal interactions or (2) both linear and nonlinear causal interactions. If the MVAR model-based DI estimate is comparable in value to data-driven DI estimate, then the interaction is predominantly linear. This is not feasible with existing metrics because they can be split into two non-overlapping groups - the first group only detects linear causal interactions (e.g., Granger causality, partial directed coherence), whereas the second group detects both linear and nonlinear causal interactions (e.g., transfer entropy). The DI estimators proposed in this chapter can be automatically adapted to other types of electrophysiological data like EEG to learn the causal connectivity.

Data-driven DI estimator seems to be more appropriate than model-based DI estimator if the underlying data distribution is not known, which is the case with most real data. The main challenge with data-driven DI estimator is estimating the causal conditional likelihood nonparametrically and its computational complexity. We used kernel density estimators in this chapter to estimate causal conditional likelihood. Kernel density estimators are asymptotically optimal [5]. Their bias decreases with increasing number of data samples and complexity increases with the dimensionality of the data, just like other nonparametric estimators. Even though we selected op-

timal bandwidth using smoothed cross-validation to minimize the asymptotic mean integrated squared error, several other criteria could also be used [5, 75]. In addition, data-driven entropy estimators based on adaptive partitioning, nearest neighbors and m-spacing algorithms [58, 89] can also be used to estimate DI nonparametrically. Another approach to estimate DI nonparametrically is to extend the universal DI estimator proposed for discrete-valued signals in [88] to continuous-valued ECoG signals. Future work should also include developing approximate data-driven DI estimators to further reduce computational complexity. We use both the MVAR model-based and data-driven DI estimators to infer the causal connectivity graph from ECoG data to identify the seizure onset zone and learn the spatiotemporal evolution of seizures in epileptic patients in the following chapter.

Chapter 4

Application to Epilepsy

4.1 Introduction

Epilepsy is a common neurological disease affecting nearly 1% of the world's population. Epilepsy is characterized by unprovoked seizures, which are periods of hypersynchronous activity in the brain. The current treatment options include medication, resective surgery and more recently, electrical stimulation approaches like vagus nerve and responsive neurostimulation. However, medication is not able to stop seizures in about one-third of the patients. The efficacy of the other current neuromodulation approaches is variable and almost never results in a cure [21, 22]. The current approaches lack specificity and suffer from negative side effects ([27] and references therein). Selective modulation of the epileptic circuits in the brain via electrical stimulation [26], optogenetics and designer receptive technologies [27] represent possible options for better treatments for this disabling disease. Our limited understanding of seizure generating mechanisms is a major bottleneck to develop better treatments for epilepsy. In this thesis, we utilize the novel information-theoretic metrics described in chapters 3 and 2 to infer the seizure generating mechanisms from electrocorticographic (ECoG) data recorded from patients with epilepsy. Specifically, directed information is used to learn the causal connectivity between various brain regions to identify seizure onset zone and to learn the spatiotemporal evolution of seizure activity. Mutual information in frequency is used to learn the cross-frequency coupling

between electrodes in seizure onset zone and infer the spectral oscillations involved in the generation of seizures.

Effective or causal connectivity [41] quantifies how the activity spreads between different brain regions and can be used to characterize epileptic networks. In addition, causal connectivity can also be used to identify seizure onset zone (SOZ) (brain regions initiating seizures [28]) and has been shown to predict the efficacy of resective surgery [50, 51]. The DI metric with model-based and data-driven estimators described in chapter 3 allows us the flexibility to simultaneously use both these estimators and identify which one leads to more reliable causal connectivity graphs from real ECoG data. This would also allows us to examine the appropriateness of imposing linear MVAR assumption on ECoG data. We used the model-based DI estimator with MVAR model assumption to detect the linear causal interactions and the data-driven DI estimator to detect both linear and nonlinear causal interactions between ECoG channels. We observed that nonlinear causal interactions between channels are stronger around the onset of a seizure, as widely believed [81].

We then describe a model-based and a data-driven SOZ identification algorithm to identify SOZ from the causal connectivity graphs inferred using model-based and data-driven DI estimators respectively. The SOZ identified by model-based and data-driven algorithms are respectively the isolated nodes and strong sources in the corresponding causal connectivity graphs. Despite the numerous SOZ identification algorithms available [46, 49–51, 110], the current clinical gold standard is still the visual analysis of ECoG data by the neurologist. We therefore compared the performance of both model-based and data-driven SOZ identification algorithms with visual analysis by the neurologist. We find that the data-driven approach outperforms the model-based approach and also leads to more interpretable results.

The ECoG recordings are analyzed during preictal, ictal and postictal periods to learn the evolution of seizure mechanisms over time. Causal connectivity is inferred from multiple sliding windows during preictal, ictal and postictal periods using both MVAR model-based and data-driven DI estimators. SOZ is observed to be less strongly connected than non-SOZ regions during seizures when MVAR model-based DI estimator is used, a counterintuitive observation. On the other hand, SOZ acts as strong source of information during preictal and ictal periods and a sink of information during postictal periods, as expected, when data-driven DI estimator is used. This essentially implies that the ECoG data is non-linear during seizures and the MVAR model-based approach is unable to capture the nonlinear dynamics, whereas the data-driven approach successfully captures the non-linear interactions in data.

Finally, the characteristics of the cross-frequency coupling (CFC) in the seizure onset zone (spatial regions from which seizures originate [28]) of epileptic patients is analyzed using the MI-in-frequency metric. While cross-frequency coupling is recently used to detect seizures [111] and delineate seizure onset zone [112, 113], we estimate CFC to investigate the dynamics of cross-frequency coupling over the duration of a seizure and beyond. We analyze the electrocorticographic (ECoG) recordings from the seizure onset zone (SOZ) channels during eleven seizures in four medial temporal lobe epilepsy patients. We observe that coupling or synchronization in gamma and ripple high frequency oscillations (HFOs) increases during seizures, when compared with preictal (immediately before the seizure) and postictal (immediately after seizure) periods. This increase is largest within a SOZ channel and almost non-existent between distinct anatomical regions in SOZ, suggesting that different regions in SOZ potentially drive the rest of brain to a seizure state independently. In addition, postictal state is characterized by a relative increase in low-frequency coupling and an

Table 4.1 : Clinical Details of the Patients Analyzed.

Patient ID	Age/Sex	Syndrome	Seizure Type	Electrode Type	Surgery	Outcome of Surgery
P1	20/M	Nonlesional temporal	CPS	D	Right TL	Class I
P2	60/M	Lesional temporal	CPS	D	Selective Left HC	Class II
P3	29/M	Nonlesional temporal	CPS	G+D	Right TL	Class II
P4	37/M	Nonlesional extratemporal	SPS+CPS	G	Right OC	Class III
P5	20/F	Lesional temporal	CPS	G+D	Left TL	Class I

CPS - complex partial seizures, SPS - simple partial seizures. D - depth electrodes, G - subdural grid electrodes. TL - temporal lobectomy, HC - hippocampectomy, OC - occipital corticosectomy. The outcomes are in Engel epilepsy surgery outcome scale. "Class I - free of disabling seizures, class II - Almost seizure-free, class III - worthwhile improvement, class IV - no worthwhile improvement" [114].

increase in linear interactions between SOZ channels.

4.2 Clinical ECoG Data

The five patients analyzed here were all managed and treated by our physician coauthors. The clinical details of these patients are summarized in Table 4.1. Three seizure records each from patients P1, P2 and P5, two from patient P3 and one from patient P4 were analyzed. Each seizure record was approximately 10 minutes long and contained one seizure. Each seizure on average lasted for a minute and was roughly in the middle of the seizure record. The seizure start and end time was identified by the neurologist. Each electrode records the voltage waveform at a sampling frequency of 1 KHz. The number of electrodes in these five patients varied from 120 to 150. Electrodes with artifacts likely due to either loose contacts, patient movement

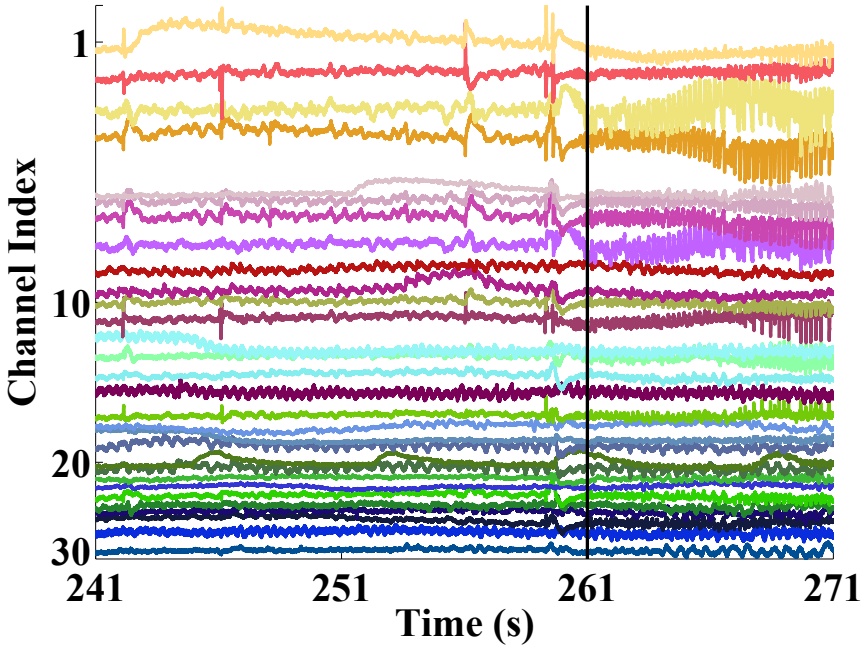


Figure 4.1 : A 30s snapshot of ECoG signals from the 30 high energy channels of P1. The seizure start time, represented by a vertical solid black line, is identified by neurologist. Causal connectivity is estimated from this entire 30s window for this seizure record.

or excessive line noise were not included in the analysis. A snapshot of ECoG data from patient P1 is plotted in Fig. 4.1.

4.3 Seizure Onset Zone Identification Algorithms

Seizure onset zone (SOZ) is defined as the regions of the brain that initiate seizures [28]. The current clinical standard is for neurologists to identify SOZ from visual analysis of the ECoG data. The SOZ identified in this way is removed during resective surgery. However, visual analysis is time consuming, subjective and potentially unreliable [49, 115]. We propose two computationally derived SOZ identification algorithms - model-based and data-driven SOZ identification algorithms. We identified the SOZ in five patients with epilepsy using these two algorithms and compared their

performance with visual analysis by the neurologist.

The first stage of the proposed SOZ identification algorithm is an energy detector which selects only M channels out of all ECoG channels for further analysis. The main objective of this stage is to reduce the computational complexity of the proposed algorithms. The energy is l_2-norm of the ECoG signal computed from a window around the start of seizures containing preictal and ictal recordings. Any channel involved in seizure onset is expected to have interictal spikes before the seizure starts and/or have high amplitude low-frequency ictal activity once the seizure is fully developed, both of which will increase the energy in the selected time-window. The time-window was selected to be long enough to capture both spiking and large ECoG amplitudes during seizures. The second stage consisted of estimating the causal connectivity between every pair of M channels selected in the first stage to form a $M \times M$ causal connectivity matrix. The causal connectivity was estimated from a shorter time-window around the seizure start time, since we are interested in estimating the seizure onset electrodes. The following subsections describe the remaining stages of the two proposed SOZ identification algorithms.

4.3.1 Model-based SOZ Identification Algorithm

In this approach, ECoG data is assumed to be derived from a MVAR process with Gaussian white noise. This is a very common assumption imposed to estimate causal connectivity between ECoG data [50, 52]. The MVAR model-based DI estimator is used to infer the causal connectivity between the selected M high energy channels. The causal connectivity estimated using this approach only represents the linear causal interactions between the ECoG channels. However it is widely believed that seizures are highly non-linear phenomenon during which SOZ drives the rest of the

network into a hypersynchronous state [21, 28, 81]. As a result, we expect the seizure onset electrodes in the causal connectivity graph to be isolated, since model-based approach can only capture linear causal interactions. The proposed model-based algorithm therefore identifies the nodes in the causal connectivity graph with zero degree (threshold was set to select only the strongest 10% connections) as the estimated SOZ. If a patient had multiple seizures, the electrodes identified across all seizures in that patient form the estimated SOZ for that patient.

4.3.2 Data-driven SOZ Identification Algorithm

In this algorithm, no parametric model assumptions were imposed on ECoG data. The causal connectivity between the M high energy channels selected in the first stage was inferred using the data-driven DI estimator. This estimator inferred both linear and nonlinear causal interactions between channels. Intuitively, activity at the SOZ electrodes drives the activity at the other electrodes into a hypersynchronous state via linear and nonlinear causal interactions [81]. We therefore expect the SOZ electrodes to act as sources (with strong outgoing and weak incoming causal connections) in the causal connectivity graph inferred around the seizure start time using data-driven DI. As a result, the SOZ nodes in the causal connectivity graph are expected to have large net-outward flow of information. The data-driven SOZ identification algorithm quantifies this intuition to estimate SOZ. The net-outward flow (Φ) of causal information from an electrode i is calculated using

$$\Phi(i) = \sum_{j=1, j\neq i}^{M} \{I(i \to j) - I(j \to i)\}. \tag{4.1}$$

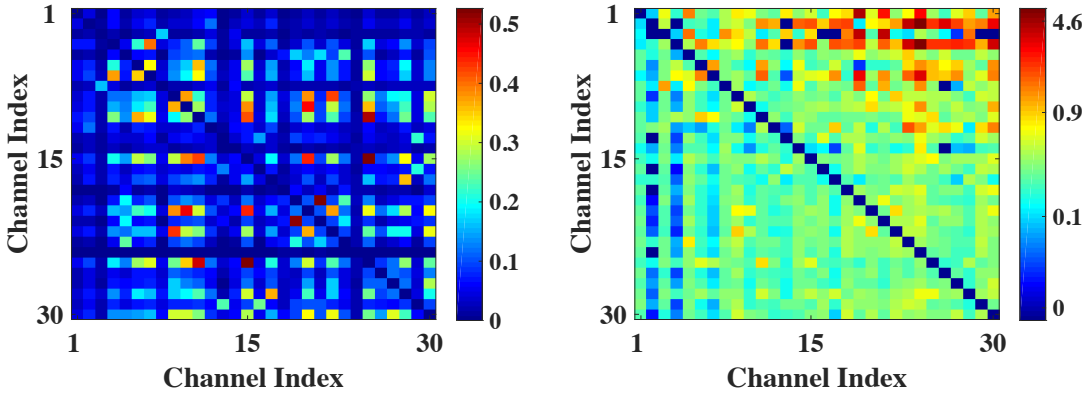
If a patient had multiple seizures, the net outward flow of an electrode is the average net outward flow of that electrode across all seizures recorded in that patient. Then the normalized net outward flow ($\tilde{\Phi}$) of the electrode i is given by

$$\tilde{\Phi}(i) = 100 \times \frac{\Phi(i)}{\sum\limits_{j:\Phi(j)>0} \Phi(j)}. \tag{4.2}$$

The electrodes with $\tilde{\Phi} > 5\%$ are considered to have significant net outward flow of information in the causal connectivity graph and are identified as the seizure onset electrodes for that patient by the data-driven SOZ identification algorithm.

4.3.3 Performance of the Proposed SOZ Identification Algorithms

The energy detector selected the top $M = 30$ channels with the largest energy computed from a 100s window comprising of 50s of activity immediately before and after the seizure starts. The causal connectivity graph between these high energy channels is then estimated using model-based and data-driven DI estimators from a 30s window that begins 20s before the seizure start time and ends $10s$ into the start of the seizure. We assumed that the current activity at an ECoG channel does not depend on more than 150ms of past activity (150 past samples at $F_s = 1$KHz) at this channel and other channels. This corresponds to restricting the model order J_{yy}, K_{yx} search space to $[1, 150]$ for the MVAR model-based DI estimator. In addition, we need to capture the connectivity just before and just after a seizure starts to estimate the SOZ. Therefore, we used ECoG data from a $30s$ window (3×10^4 data points) that begins $20s$ before the start of the seizure to be stationary. The same window was used for the data-driven estimator as well. In addition, the past activity was down-sampled by a factor of 50 for the data-driven estimator to restrict the J_{yy}, K_{yx} search space to



(a) From model-based algorithm (b) From data-driven algorithm

Figure 4.2 : Causal connectivity between 30 high energy channels estimated from ECoG data between 241s and 271s from the second seizure of P1. The channel indices with bluish rows and bluish columns (correspond to low DI estimates) in Fig. 4.2a correspond to isolated nodes and are the estimated SOZ using model-based algorithm. The corresponding channels in Fig. 4.2b have large net-outflows of information and are the estimated SOZ from data-driven algorithm.

[1, 4] and also reduce its computational complexity (i.e. the past activity of channel X can include $\{x_n, x_{n-50}, x_{n-100}, x_{n-150}\}$). The exact values of these parameters is not crucial as the algorithms seem to be fairly robust to changes in these parameters.

Consider the second seizure record of patient P1. The energy detector selected 30 high energy channels. Fig. 4.1 shows the recordings from these channels in the 30s window in which causal connectivity graph is inferred. The inferred graph by model-based and data-driven approaches is shown in Fig. 4.2. The weighted adjacency matrix of the inferred causal connectivity graphs, whose $(i, j)^{th}$ element is the DI estimate from channel i to j for $i, j \in [1, 30]$, is plotted in Fig. 4.2 using a image plot. It is clear from this figure that the mean strength of the DI estimates using model-based approach is smaller than using data-driven approach (colorbar ranges are different in the two sub-figures). We observed this across all the remaining seizures as well. This is evident from Fig. 4.3, where the mean value of DI estimate across all

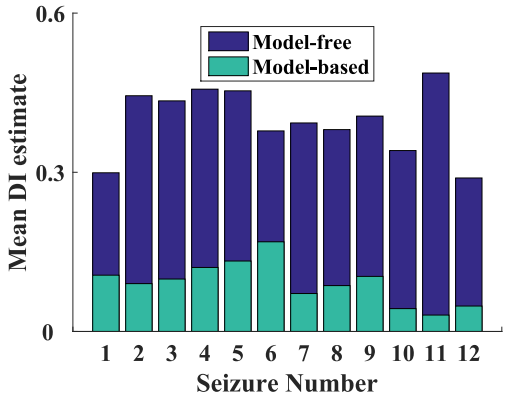


Figure 4.3 : Mean value of DI estimates obtained using model-free and model-based DI estimators from the twelve seizures in five patients with epilepsy analyzed.

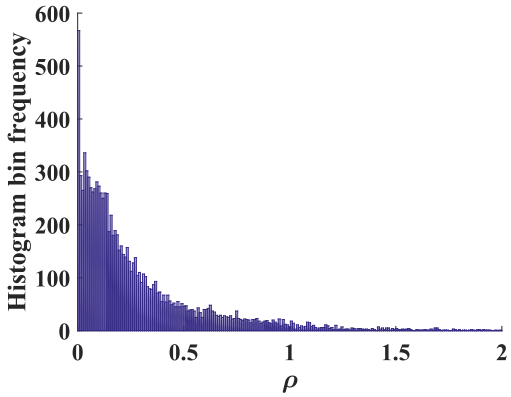
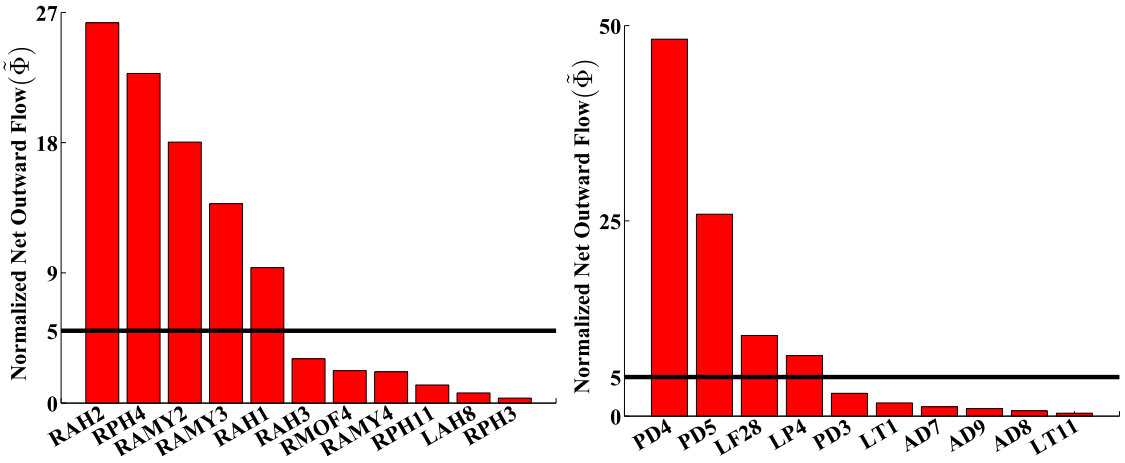


Figure 4.4 : Histogram of the ratio ρ of model-based DI estimate with model-free DI estimate between all pairs of channels from the twelve seizures in five patients with epilepsy analyzed.

channel pairs computed from model-free and model-based DI estimators is plotted. We define ρ as the ratio of model-based DI estimate and model-free DI estimate. When $\rho < 1$, model-based estimate is smaller than model-free DI estimate. The histogram of ρ between all pairs of channels in the twelve seizures analyzed is plotted in Fig. 4.4. It is clear from Fig. 4.3, Fig. 4.4 that model-based DI estimate is smaller than model-free DI estimate in most cases. This implies model-free approach captures more causal information than model-based approach, because it can capture nonlinear interactions.

The nodes with zero degree in the causal connectivity graphs from each seizure in a patient are identified as the SOZ by the model-based algorithm. The zero degree criterion used by model-based algorithm is counterintuitive, since we except the SOZ to drive the network to seizure state and not be weakly connected. On the other hand, the data-driven algorithm selects electrodes with large net outflows, which is very intuitive. The data-driven algorithm computed the normalized net outward flow



(a) From Patient P1 (b) From Patient P3

Figure 4.5 : Normalized net outward flow from the ECoG electrodes with positive net information outflow using data-driven SOZ identification algorithm.

Table 4.2 : Seizure onset zone identified from the proposed algorithms and the visual analysis by neurologist.

Patient - # of Seizures	**Model-based Algorithm**	**Data-driven Algorithm**	**Visual Analysis**
P1 - 3	*RAH 1-3*, *RPH 2-4*	*RAH 1-2*, *RPH 4*, RAMY 2-3	RAH 1-3, RPH 2-4, RAMY 2-3
P2 - 3	*LAH 2-4*, *LPH 1-2*	*LAH 2-4*, *LPH 2*	LAH 2-4, LPH 1-2
P3 - 2	LT 1-3, 10	*PD 4-5*, LF 28, LP 4	PD 3-5
P4 - 1	*LO 3, 14, 15, 25*, LO 12, 13, *PST 3*, PST 1, MOG 27	*LO 3, 14, 15*, 12, PST 1, MOG 23, SOG 21, 36	LO 3, 14, 15, LO 25, PST 3
P5 - 3	*MST 1, 2*, *HD 1*	*MST 1, TP 1*, *HD 1*	MST 1, 2, TP 1, HD 1-3, AST 2

The label of an ECoG electrode comprises of an abbreviation of the brain region it is implanted in and a number. For depth electrodes, smallest number is assigned to deepest electrode from scalp. For instance, RAH1 - deepest electrode contact in depth electrode in right anterior hippocampus and LO3 - third electrode contact in subdural grid electrode over lateral occipital lobe. RPH - right posterior hippocampus, RAMY - right amygdala, LF - lateral frontal, LP - lateral parietal, LT - lateral temporal, PD - posterior hippocampal depth, MOG - medial occipital grid, SOG - sub-occipital grid, PST - posterior sub-temporal, MST - mid-subtemporal lobe AST - anterior sub-temporal lobe, TP - temporo-polar, HD - hippocampal depth.

for each node using (4.2). Fig. 4.5a plots the $\tilde{\Phi}$ for all electrodes with positive net outward flows in patient P1. The electrodes with $\tilde{\Phi} > 5\%$ are the estimated SOZ for this patient P1 using data-driven algorithm.

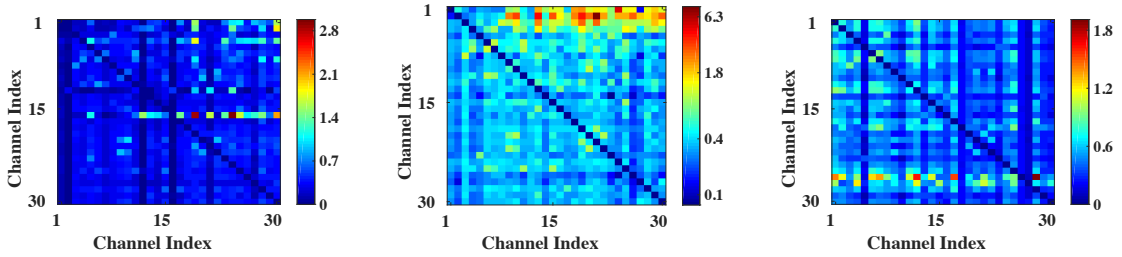
Table 4.2 summarizes the results from our analysis. The first column in Table 4.2 identifies the patient ID and the number of seizures analyzed for that patient. The second, third and fourth columns in Table 4.2 list the SOZ identified across all the five patients using model-based, data-driven algorithms and visual analysis respectively. We observed that all the channels identified as SOZ by visual analysis, except AST 2 in one seizure of P5, are included in the 30 high energy channels selected from each seizure by the energy detector in the first stage. The top 30 channels selected from two seizures in patient P5 contained AST 2, but the 30 channels picked from the third seizure did not contain AST 2. The normalized net outflow $\tilde{\Phi}$ from AST 2 electrode for patient P5 using data-driven algorithm was 1% and hence this electrode was not identified as SOZ (note that $\tilde{\Phi}$ has to exceed 5% to be selected as SOZ). Expect for this one region, it is clear from this table that the data-driven algorithm identifies all the regions identified by the neurologist, whereas the model-based algorithm misses some regions (for instance, RAMY electrodes in P1, TP and AST electrodes in P5). Also, the model-based algorithm incorrectly identified lateral temporal (LT) electrodes as SOZ in patient P3, whereas data-driven algorithm correctly identified posterior depth (PD) electrodes in hippocampus as SOZ. Except in P3 and P4, both algorithms do not have any false positives. The false positives in P4 could be because only one seizure was analyzed in this patient.

Another advantage of the data-driven SOZ identification algorithm over model-based algorithm and analysis by the neurologist is that $\tilde{\Phi}$ could be used as a quantitative metric to rank the electrodes in the decreasing order of clinical relevance. Fig. 4.5

plots the $\tilde{\Phi}$ of all electrodes with positive net outward flows in patients P1 and P3. Using our metric $\tilde{\Phi}$, it is clear from Fig. 4.5b that electrodes PD4, PD5 contribute much more in generating and spreading seizures than LF28 and LP4 electrodes even though $\tilde{\Phi}$ exceeds the chosen threshold at all these four electrodes. Depending on the significance level (5% is used here), the set of selected SOZ electrodes varies. We observed in all five patients that the electrodes with the highest $\tilde{\Phi}$ values were always the same as the ones identified by the neurologist. Visual analysis by the neurologist can only give qualitative information about the SOZ and cannot give quantitative information like the proposed data-driven SOZ identification algorithm. In addition, data-driven algorithm can also differentiate between electrodes in close proximity - for example, $\tilde{\Phi}$ is negative for RPH2 electrode in P1 even though $\tilde{\Phi}$ is positive for both RPH3 and RPH4 (refer to Fig. 4.5a). The increased spatial-specificity provided by our data-driven algorithm could be relevant for next generation epilepsy treatments [27]. The main advantage of the model-based algorithm over data-driven one is its lower computational complexity. However, this is less critical with today's powerful computers. To summarize, data-driven SOZ identification algorithm outperforms model-based algorithm and provides more interpretable results.

4.4 Spatiotemporal Evolution of Seizures

The causal connectivity estimated from a single short time window is used to identify the seizure onset zone in the preceding section. In this section, we extend the analysis to estimate the causal connectivity over time (referred to as dynamic causal connectivity). Causal connectivity is estimated from a time window and then the windows are shifted in time to learn the dynamic causal connectivity. The reason for this analysis is to understand the differences in causal connectivity between different regions



(a) From segment before seizure (b) From segment during seizure (c) From segment after seizure

Figure 4.6 : Causal connectivity between the 30 high energy channels from the second seizure of patient P1 estimated using data-driven DI estimator from ECoG data in three segments - one before seizure (181s -211s), one during seizure (261s - 291s) and one after seizure (361s - 391s). This seizure starts at 261s and ends at 350s.

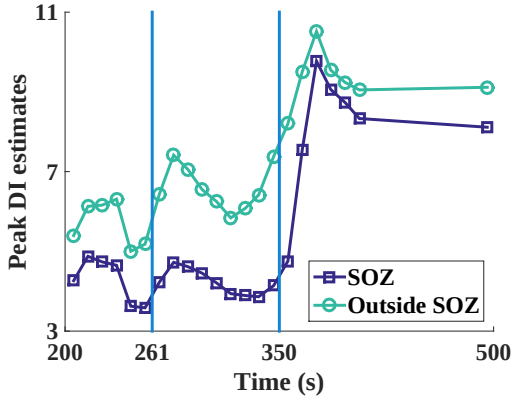


Figure 4.7 : Average values of the peak DI estimates from an electrode in SOZ and an electrode outside SOZ obtained using MVAR model-based DI estimator over the duration of a seizure. The vertical lines at 261s and 350s correspond to seizure start and end times respectively.

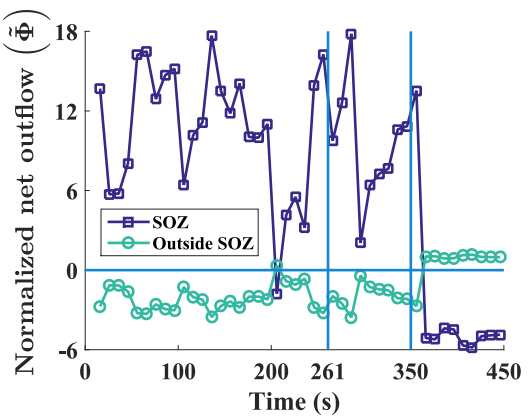


Figure 4.8 : Average normalized net outflow $\tilde{\Phi}$ from an electrode in SOZ and an electrode outside SOZ using data-driven DI estimator over the duration of a seizure. The vertical lines at 261s and 350s correspond to seizure start and end times respectively.

during seizures when compared to the period just before seizures. This analysis can also be potentially used to hypothesize how seizures are generated.

We used the MVAR model-based and data-driven DI estimators to estimate causal connectivity between the ECoG channels from multiple windows in preictal, ictal and

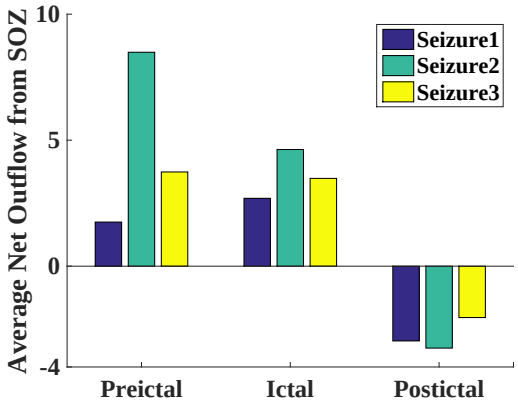


Figure 4.9 : Average normalized net outflow $\tilde{\Phi}$ from an electrode in SOZ obtained using data-driven DI estimator during the preictal, ictal and postical periods from the three seizures analyzed in patient P1.

postictal periods. The causal connectivity before, during and after the second seizure of patient P1 estimated using data-driven DI estimator from three 30s long windows is shown in Fig. 4.6. It is clear from Fig. 4.6 that SOZ electrodes (corresponding to rows with more red color or large DI values in Fig. 4.6b) have large net outflows during seizure when compared with before and after seizure (same rows have more blue color or smaller DI value in Fig. 4.6a , 4.6c). The average value of the peak DI estimates from all channels in SOZ and outside SOZ obtained from MVAR model-based DI algorithm is plotted in Fig. 4.7, whereas the mean of the normalized net outward flow $\tilde{\Phi}$ from all electrodes in SOZ and from all electrodes outside SOZ versus time is plotted in Fig. 4.8. It is clear from this figure that SOZ electrodes are weakly connected during seizures when MVAR model-based approach is used, a result which is consistent with what is typically reported in the scientific literature [53, 116]. On the other hand, tt is clear from Fig. 4.8 that SOZ electrodes have large positive $\tilde{\Phi}$ (sources) even before seizure is clinically manifested and have negative $\tilde{\Phi}$ (sinks) once the seizure ends. This suggests SOZ continuously tries to drive the rest of the

brain into a seizure and becomes deactivated as soon as the seizure ends. To test this hypothesis further, we extended the analysis using data-driven DI estimator to all the three seizures in patient P1 and the results are plotted in Fig. 4.9. It is clear from this figure that, SOZ acts as strong sources of causal information before a seizure starts (preictal) and during seizures (ictal), whereas they act as sinks of causal information once the seizure ends (postictal), consistent with our intuition. This once again suggests that ECoG recordings around seizure are highly nonlinear and that the analysis should be done using data-driven approaches which capture nonlinear interactions and not using MVAR model-based approaches.

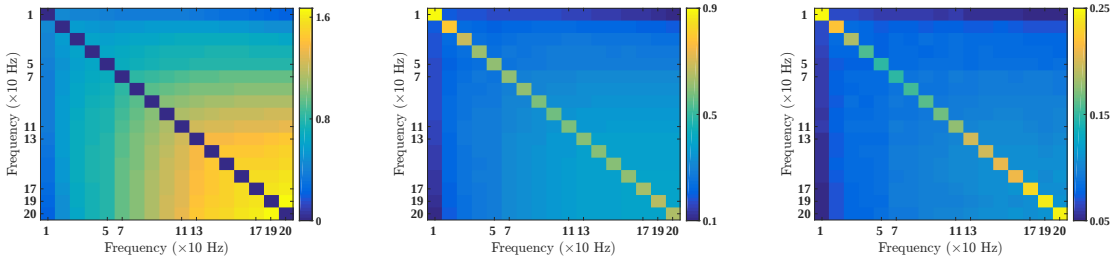
4.5 Cross-Frequency Coupling in Seizure Onset Zone

We used our newly defined metric, MI-in-frequency, to infer the coupling across frequency in the ECoG recordings from seizure onset zone (SOZ). The goal is to determine whether the oscillations in alpha (7.5-12.5 Hz), beta (12.5 - 30 Hz), gamma (30-80 Hz) and ripples (80-200 Hz) are independent or not and if so, quantify their dependence using MI-in-frequency.

We analyzed ECoG data, sampled at $F_s = 1$ KHz, from eleven seizures in four patients, P1, P2, P3 and P5, with medial temporal lobe epilepsy. Clinical details of the patients are summarized in Table 4.1. The seizure start and end time, along with the seizure onset zone was marked by the neurologist. The ECoG recordings from SOZ electrodes were analyzed during preictal (window spanning 3 minutes immediately before seizure starts), ictal (during seizures) and postictal (window spanning 3 minutes immediately after seizure ends) periods. The number of data samples from each SOZ electrode during the preictal and postictal periods is 180×10^3 (smaller in cases when the 3 minutes of recordings were not available) and during ictal period

is dependent on the duration of the seizure. We set $N_f = 100$, which implies the spectral process increments can be estimated at integral multiples of 10 Hz, up to 500 Hz. As mentioned earlier, we only focus on the oscillations up to 200 Hz, excluding the harmonics of 60 HZ due to line noise, and estimate 17×17 MI-in-frequency using nearest neighbor based estimator (section 2.4.2) during preictal, ictal and postictal periods in all the eleven seizures.

We focus on the average CFC within each ECoG electrode in SOZ, between two electrodes in the same anatomical region in SOZ and between electrodes in different anatomical regions in SOZ during preictal, ical and postictal periods. For instance, SOZ in patient P1 comprises of 8 electrodes in 3 different anatomical regions–RAH, RPH and RAMY (Table 4.1). For this patient, we estimate 8 MI-in-frequency matrices, one per SOZ electrode, to learn the average CFC within an electrode in SOZ. We estimate 14 MI-in-frequency matrices (6 to learn the CFC between the 3 SOZ electrodes in RAH, 6 to learn the CFC between the 3 SOZ electrodes in RPH region, 2 to learn the CFC between the 2 SOZ electrodes in RAMY region) to learn the average CFC within an anatomical region in SOZ. We estimate the remaining 42 MI-in-frequency matrices (56 to learn the CFC every pair of two electrodes in the SOZ, minus the 14 required to learn the CFC between electrodes in the same region) to learn the CFC between different regions in SOZ. All the resulting 17×17 MI-in-frequency matrices are averaged across all patients and plotted in Fig. 4.10, Fig. 4.11 and Fig. 4.12. Also note that even though MI-in-frequency between two different electrodes is not symmetric, it is symmetric in plots like Fig. 4.10b and Fig. 4.10c because we averaged MI-in-frequency matrices between all possible permutations of the relevant electrodes in SOZ. Statistical significance of the MI-in-frequency estimates at each frequency pair was individually assessed using $N_p = 10$ permuted estimates,

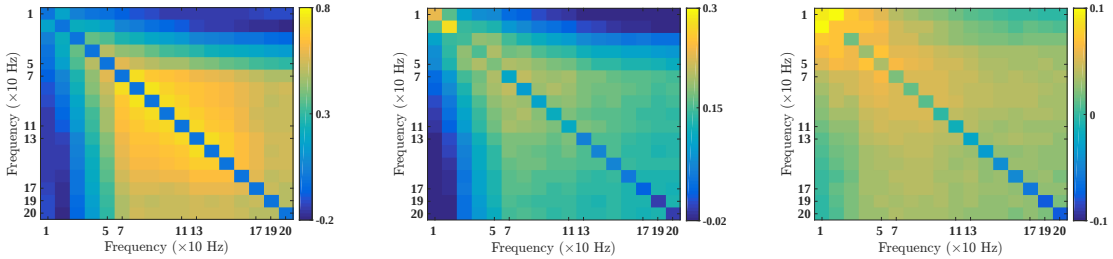


(a) Within each electrode in SOZ (b) Between electrodes in same region in SOZ (c) Between electrodes in different regions in SOZ

Figure 4.10 : Cross-frequency coupling in the preictal period in seizure onset zone. In Fig. 4.10a, estimates of MI-in-frequency over the frequencies $\{10, 20, \cdots, 200\}$ Hz excluding $\{60, 120, 180\}$ Hz are obtained from each electrode in SOZ and the resulting CFC estimates are averaged over all SOZ electrodes in the eleven seizures from the four temporal lobe epilepsy patients and plotted. Similarly, in Fig. 4.10b, MI-in-frequency is estimated between the various frequency components in two ECoG electrodes that are in the same anatomical region and the resulting average is plotted. For instance, in patient P1, ECoG electrodes RAH1, RAH2 are in the same anatomical region, whereas RAH1, RPH1 are in different anatomical regions. In Fig. 4.10c, MI-in-frequency is estimated between two ECoG electrodes in different anatomical regions and the resulting average CFC is plotted.

according to the procedure described in section 2.4.3, across all frequency pairs, the spatial parameters (within each electrode, between electrodes in same region and between electrodes in different regions) and temporal parameters (preictal, ictal and postictal) considered. If there is no statistically significant MI-in-frequency between them, it is set to zero. The CFC matrices with statistically significant MI-in-frequency estimates are then averaged across the eleven seizures considered for the 3 spatial and 3 temporal parameters, resulting in 9 averaged MI-in-frequency matrices (which are analyzed in Fig. 4.10, 4.11 and 4.12).

Fig. 4.10a plots the averaged MI-in-frequency coupling matrix during preictal period. The $(i, j)^{\text{th}}$ element in the matrix in Fig. 4.10a is the averaged MI-in-frequency between the $10i$ and $10j$ Hz frequency components during preictal period across all



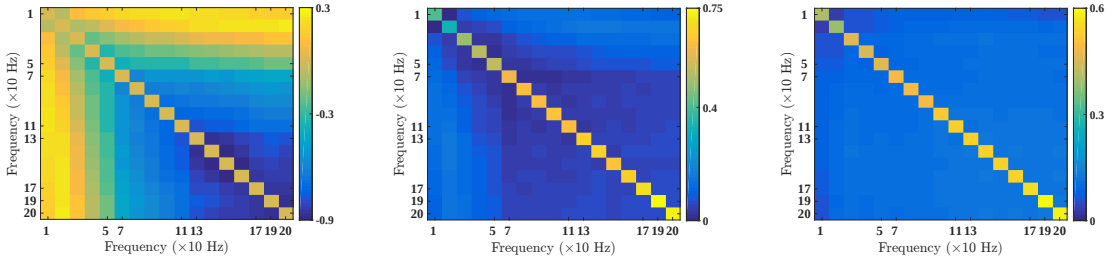
(a) Within each electrode in SOZ (b) Between electrodes in same region in SOZ (c) Between electrodes in different regions in SOZ

Figure 4.11 : Difference in the cross-frequency coupling between ictal and preictal periods in seizure onset zone. In Fig. 4.11a, estimates of MI-in-frequency over the frequencies $\{10, 20, \cdots, 200\}$ Hz excluding $\{60, 120, 180\}$ Hz are obtained from each electrode in SOZ in ictal period and the difference between the averaged CFC estimate in ictal and preictal period (which is shown in Fig. 4.10a) is plotted. Similarly, in Fig. 4.11b, MI-in-frequency is estimated between the various frequency components in two ECoG electrodes that are in the same anatomical region in ictal period and the difference in averaged CFC estimate between ictal and preictal period is plotted. In Fig. 4.11c, MI-in-frequency is estimated between two ECoG electrodes in different anatomical regions and the difference in averaged CFC estimate between ictal and preictal period is plotted.

SOZ electrodes in the eleven seizures analyzed. The average CFC between two SOZ electrodes in the same anatomical region and in different anatomical regions in the preictal period in respectively plotted in Fig. 4.10b, Fig. 4.10c. It is clear from this figure, coupling across frequency within each SOZ electrode is larger than between electrodes in same anatomical regions, which in turn is bigger than between different regions in SOZ. However, the linear component or equivalently, the CFC along the diagonal is relatively high both within and across regions in SOZ. This suggests that the neighboring regions in SOZ have relatively strong linear interactions (possibly due to the close distance between them) just before a seizure starts. In addition, ripples are heavily synchronized during preictal stage, when compared with the low-frequency oscillations.

The difference of the averaged MI-in-frequency coupling matrix between ictal and preictal periods is plotted in Fig. 4.11. The average difference of CFC within each SOZ electrode, between electrodes in the same region and between different SOZ regions is plotted in Fig. 4.11a, Fig. 4.11b and Fig. 4.11c respectively. First from Fig. 4.11a, the synchronization between all frequency pairs, particularly in gamma and ripples, seems to increase in a SOZ electrode when compared to just before seizure. This effect is also accompanied by a small decrease in low-frequency coupling. The increase is high frequency coupling is relatively smaller within a SOZ region (Fig. 4.11b). Another interesting observation is the lack of increase in the linear coupling between electrodes in SOZ (referring to the diagonal in Fig. 4.11b, Fig. 4.11c), which suggests that seizures are not accompanied by an increase in linear coupling, but rather by an increase in nonlinear interactions between electrodes. The real surprise finding, however, is the very small increase in coupling between different SOZ regions during seizures when compared to preictal periods (Fig. 4.11c). This lack of coupling was most severe in patient P1, where causal connectivity analysis in [20] suggested that there is no clear dominant region among the three regions (RAH, RPH and RAMY) in SOZ (refer to Fig.9a in [20]). This suggests that epilepsy is patient-specific and different SOZ regions can potentially drive the rest of the brain into a seizure state independently, which implies any non-surgical treatment should target these different regions simultaneously to disrupt the epileptic network.

Finally, the difference in CFC estimates between postictal and ictal periods within each SOZ electrode, between electrodes in same region and across regions in SOZ is plotted in Fig. 4.12a, Fig. 4.12b and Fig. 4.12c respectively. The synchronization in high frequency bands decreases and low frequencies become more synchronized in postictal period compared to ictal period within SOZ electrode. The major change



(a) Within each electrode in SOZ (b) Between electrodes in same region in SOZ (c) Between electrodes in different regions in SOZ

Figure 4.12 : Difference in the cross-frequency coupling between postictal and ictal periods in seizure onset zone. In Fig. 4.12a, estimates of MI-in-frequency over the frequencies $\{10, 20, \cdots, 200\}$ Hz excluding $\{60, 120, 180\}$ Hz are obtained from each electrode in SOZ in postictal period and the difference between the averaged CFC estimate in postictal and ictal period is plotted. Similarly, in Fig. 4.12b, MI-in-frequency is estimated between the various frequency components in two ECoG electrodes that are in the same anatomical region in postictal period and the difference in averaged CFC estimate between postictal and ictal period is plotted. In Fig. 4.12c, MI-in-frequency is estimated between two ECoG electrodes in different anatomical regions and the difference in CFC estimate between postictal and ictal period is plotted.

in coupling between electrodes is the increase in the linear coupling, which suggests that postictal periods, unlike ictal periods, are characterized by an increase in linear interactions. These results highlight the role of gamma and ripple high frequency oscillations (HFOs) during seizures and the dynamic reorganization of synchronization between neuronal oscillations inside the seizure onset zone during the course of a seizure.

These results highlight the role of gamma and ripple high frequency oscillations (HFOs) during seizures and the dynamic reorganization of synchronization between neuronal oscillations during the course of a seizure in SOZ channels. In addition, SOZ channels also seem to independently drive the rest of the brain during seizures, which can be verified by analyzing the CFC between channels in SOZ and outside SOZ. In addition, we need to classify whether the excess synchronization during seizures

is pathological or physiological by comparing the MI-in-frequency coupling matrices during seizures with those during interictal periods as the baseline. Analyzing the pathological oscillations occurring during seizures could improve our understanding of epileptic seizures, and potentially lead to better treatments in the future.

4.6 Discussion and Conclusions

Directed information was used in this chapter to estimate causal connectivity between ECoG channels. The causal connection identified between two channels could be due to the effect of activity at other spatial locations in the brain. If an ECoG electrode was implanted at these other locations, causally conditioned DI can be used to remove their influence. This was demonstrated using the four node examples in section 3.5. On the other hand, if ECoG activity is not recorded from these locations, then removing the effects of these hidden nodes on the inferred causal connectivity is a very hard problem in general. Future work should look into the sensitivity of DI to volume conduction effects when compared with synchronization metrics like phase lag index [117].

DI estimators proposed in this thesis do not quantify the amount of causal information between time-series at each frequency, unlike partial directed coherence (PDC) or directed transfer function (DTF) (note that mutual information in frequency only quantifies amount of statistical dependence between the data and does not find any directionality for the dependence in frequency). However, the advantage of DI is that data-driven DI estimator can detect nonlinear causal interactions, which PDC or DTF cannot detect. Metrics based on PDC, DTF assume the data is drawn from a MVAR model and can only detect linear causal interactions (similar to MVAR model-based DI estimator proposed in this paper). To demonstrate this, we

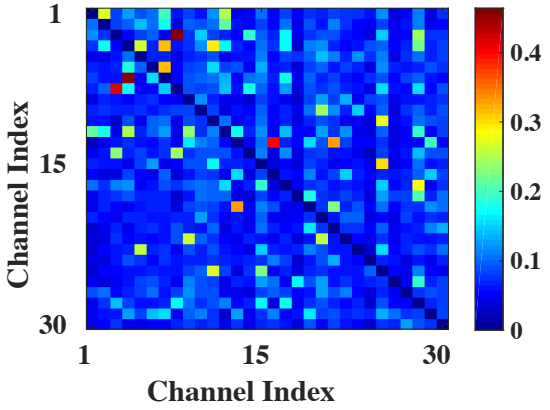


Figure 4.13 : Causal connectivity between 30 high energy channels depicted in Fig. 4.1 estimated using partial directed coherence.

estimated the causal connectivity graph between the 30 channels depicted in Fig. 4.1 by PDC using eMVAR toolbox [61]. The resultant 30×30 causal connectivity matrix is plotted in Fig. 4.13, in which (i, j) element corresponds to the maximum value of PDC from channel i to channel j. Note that causal connectivity estimates from the proposed DI estimators for the same data is plotted in Fig. 4.2. Comparing Fig. 4.13 with Fig. 4.2b, it is clear that net outflow from the SOZ electrodes is not large in PDC when compared to data-driven DI. This implies unlike data-driven DI estimator, PDC cannot capture nonlinear causal interactions.

We also proposed model-based and data-driven algorithms to identify the SOZ. The first stage of both these algorithms is an energy detector. The chosen electrodes from the first stage turned out to have large overlap (more than half) across multiple seizures within a patient. All electrodes with low rhythmic gamma activity in SOZ were selected by the energy detector in all the patients analyzed. Note that other criteria could also be used instead of energy detector. In particular, we experimented with selecting channels displaying strong high-frequency activity around the seizure

start time (since channels involved in seizure onset display strong high-frequency activity around the beginning of a seizure that typically develops into high amplitude low-frequency activity). The time-window used to estimate the high-frequency activity should be of much smaller length than the one used with energy detector, because the seizures typically display low amplitude rhythmic high-frequency oscillations only for a very short duration. The resulting performance with energy detector or the high-frequency activity detector was similar. We therefore presented the results only with the energy detector in this paper.

The causal connectivity graphs between the selected high energy channels estimated using MVAR model-based DI and data-driven DI from same time-window are not the same, since both estimators capture different causal interactions in the data - model-based captures linear interactions, whereas data-driven captures both linear and nonlinear causal interactions. Therefore the criterion used to estimate SOZ from the causal connectivity graph was different for the two algorithms. In the model-based approach, the SOZ nodes are isolated since they drive the other brain regions into a seizure through nonlinear interactions (which are not captured by model-based DI estimator). Similar results were reported in other studies using linear metrics [53, 116]. It is reported in [53] that SOZ electrodes form an isolated focus using symmetric coherence metric that captures linear interactions. On the other hand in causal connectivity graphs estimated by data-driven DI, the outgoing and incoming edges from SOZ electrodes have large and small DI estimates respectively (refer to Fig. 4.5). This is in accordance with our intuition that the SOZ drives the seizure activity [21, 28, 81]. Also metrics closely related to net outward flow were used in [46] to infer SOZ using transfer entropy (which detects nonlinear interactions) by analyzing hours of ECoG recordings (here we are only using recordings from a 30s window).

We then used the proposed DI estimators to learn the changes in causal connectivity during preictal, ictal and postictal periods. We observed that inference from the data-driven approach is in line with our intuition, suggesting that nonlinear interactions dominate the ECoG recordings around the seizure start time. This analysis should be extended to larger patient cohorts and also include interictal periods. The results from this analysis potentially could improve our understanding of spatiotemporal evolution of seizure activity and lead to the development of novel nonsurgical treatments for epilepsy.

Finally, we used the MI-in-frequency estimators to infer the coupling between neuronal oscillations before, during and after seizures in the seizure onset zone. We observe that the high-frequency synchronization within an ECoG electrode in SOZ increases during seizures and decreases immediately after seizure, which is accompanied by an increase in low-frequency coupling. Moreover, coupling between neighboring electrodes in an anatomical region in SOZ also increase during seizures when compared with preictal periods. However, the coupling between different anatomical regions in SOZ does not increase noticeably during seizures, except for a decrease in linear interactions, followed by a large increase in linear interactions immediately after a seizure. In addition, there is some variability across patients in CFC characteristics. These observations suggest that seizure activity is characterized by nonlinear interactions and is potentially due to the independent efforts by various regions within SOZ, which implies that all these regions are potential spatial targets for electrical stimulation. Going forward, the MI-in-frequency metric should be applied to infer the CFC between channels in SOZ and outside SOZ to learn how SOZ drives the rest of the brain into a seizure state in each epilepsy patient. Also, the CFC estimates during the course of a seizure should be benchmarked against interictal periods to dif-

ferentiate between physiological and pathological variation. The results from such an analysis will improve our understanding of the CFC mechanisms underlying seizure activity and will serve as the first step towards the development of patient-specific, closed-loop, non-surgical treatments for epilepsy.

Chapter 5

Conclusions and Future Directions

5.1 Innovations of the Thesis

This thesis develops novel information-theoretic metrics to solve the problem of detecting and quantifying the spectral and spatiotemporal relationships between data. The novel metrics developed are then applied to the electrocorticographic recordings from epilepsy patients to learn the characteristics of epileptic networks.

Specifically, we defined a new metric, MI-in-frequency, to quantify the statistical dependence between different frequency components in a signal, or between signals. A kernel density based and a nearest neighbor based data-driven estimator of MI-in-frequency is proposed and their performance is compared on simulated data. We observed that nearest neighbor based estimator is superior, since it is more accurate, converges faster and not as computationally intensive as kernel density based estimator. We then developed a data-driven estimator to estimate mutual information between dependent data. The main novelty of the proposed data-driven MI estimator lies in utilizing MI-in-frequency, a frequency domain metric, to estimate MI, a time-domain metric and its performance is validated on simulated data.

We developed an almost-surely convergent MVAR model-based and data-driven estimators of directed information to infer the causal connectivity graph between data recorded from multiple sensors. The performance of the proposed estimators is benchmarked on simulated data. We observed that MVAR model-based estimator

captures linear interactions between data, whereas the data-driven estimator captures both linear and nonlinear interactions. In addition, MVAR model-based DI estimator outperforms the data-driven DI estimator if the data can be modeled using a MVAR model. If not, data-driven DI estimator is superior.

The proposed DI and MI-in-frequency estimators are then applied to ECoG recordings from epilepsy patients to learn the characteristics of seizure onset zone (SOZ) around seizures. We proposed a model-based and a data-driven SOZ identification algorithm and compared their performance against visual analysis by a neurologist. We observed that data-driven SOZ identification algorithm outperforms the model-based SOZ identification algorithm. In addition, we applied the model-based and the data-driven DI estimators across multiple time-windows in the preictal, ictal and postictal periods to learn the changes in the causal connectivity over time. Using data-driven approach, we observed that SOZ acts a source during preictal and ictal periods, whereas it acts as a sink during postictal period–consistent with our intuition that SOZ drives the rest of the brain into a hyper-synchronous state. The inferences from MVAR model-based approach are not consistent with our intuition, implying that it was unable to capture the underlying nonlinear interactions between the ECoG recordings around seizure state. We then applied the MI-in-frequency metric to infer the cross-frequency coupling between seizure onset zone electrodes during preictal, ictal and postictal periods. We observed that high frequency oscillations become more synchronized during seizures, when compared with preictal and postictal periods within each SOZ channel. In addition, we only observe a small increase in cross-frequency coupling between different anatomical regions in SOZ during seizures, which implies that any potential electrical stimulation based treatment should target the different anatomical regions in SOZ simultaneously. The spectral and spatiotem-

poral analysis of ECoG data presented here is the first step towards development of effective non-surgical treatments for epilepsy.

5.2 Future Directions

Some exciting directions that are worth exploring are highlighted below.

- We developed data-driven estimators for the metrics developed in this thesis. One of the main motivations for doing so is that a good parametric model for the data is not known and it is often not linear. In domains like neuroscience, it is worth-while to develop parametric estimators of the proposed metrics when the data is modeled by specific families of nonlinear models, like a squared nonlinearities or sigmoid or deep learning based models [118].

- We focused on detecting pairwise metrics to infer the structure between a network of data streams. It is worth extending this work to infer the joint structure. A straightforward way to accomplish this is by defining conditional metrics– conditional MI-in-frequency, causally conditioned DI. However, estimating the conditional metrics becomes impossible even for small networks (say with 10 nodes) due to the curse of dimensionality. This problem is solvable if the underlying model is MVAR. However, neural data is highly nonlinear and this problem could be potentially solved in the future if the underlying model space is constrained by modeling neural data using specific families of nonlinear models.

- The novel information-theoretic metrics developed in this thesis should be applied to larger patient cohorts and across longer time periods and all ECoG channels. The data-driven DI estimator should applied to the entire ECoG

record to infer seizure mechanisms by examining the changes in causal connectivity estimated from preictal, ictal, postictal periods, when compared with interictal periods. The MI-in-frequency metric should be applied to the recordings from all electrodes outside SOZ and also between an SOZ and a non-SOZ electrode to infer the dynamics in cross-frequency coupling over the entire ECoG record. The results from these analyses have the potential to improve our understanding of seizure mechanisms and eventually lead to the development of novel nonsurgical treatments for epilepsy.

Appendix A

Online Bayesian Change Point Detection

A.1 Introduction

Epilepsy is a dynamic disease in which brain transitions between different states [40]. Each state is defined by a connectivity graph that factorizes the joint probability distribution over the activity at different electrode locations [41] and the activity is measured by electroencephalography (EEG) and electrocorticography (ECoG) recordings. The dynamic behavior observed in EEG and ECoG recordings makes the selection of optimal temporal and spatial locations for stimulation non-trivial [119]. Computational approaches for selecting the optimal electrical stimulation parameters require the complete knowledge of different states and the temporal transitions between them. As mentioned in introduction chapter of this thesis, the nonstationarity present in ECoG data can be addressed either using sliding windows or change point detection algorithms. We used sliding window approach in chapter 4. However, in this appendix, we address the problem of detecting time-points in the EEG and ECoG recordings after which the underlying state (represented by a joint probability distribution) changed, henceforth referred to as change points (CP).

The problem of detecting change points from a time series is a well-studied problem with applications in domains like finance, engineering, and medicine. Unlike the traditional solutions, any solution to the problem of segmenting epileptic activity should have low complexity, work in online instead of offline mode, and be able to

deal with the non-stationary property of EEG and ECoG data. The Bayesian CP detection algorithms presented here satisfy all these requirements. In this work, we first develop an online Bayesian change point detection algorithm that works for non independent and identically distributed (non-i.i.d.) data. This algorithm, based on the online CP detection algorithm in [120], has quadratic complexity in the number of data sample points. Secondly, an approximate algorithm based on ideas from list decoding [121] is also proposed which has linear complexity in the number of data sample points. The performance of both these algorithms is evaluated and compared with the forward/backward CP detection algorithm [122] on simulated data. Finally, the ECoG activity measured from an epileptic patient is segmented into different states using both these algorithms.

The main contributions of this work are:

- An online Bayesian change point detection algorithm for the general case of non-i.i.d. data.
- An approximate algorithm with linear complexity, based on ideas from list decoding.
- Segmenting ECoG data to identify segments of activity corresponding to the different epileptic brain states.

A.1.1 Related Work

Detecting certain events of interest like seizures and spikes from EEG and ECoG is a well-studied problem [123]. To the best of our knowledge, no one has looked into segmenting the activity in an entire observation window to find all the different states. Some of the earliest works in Bayesian change point detection are based on Markov

Chain Monte Carlo (MCMC) and its variations [124, 125]. Product partition models (PPM) for change point detection were introduced in [126]. A forward/backward algorithm to solve for CPs in data modeled by PPMs proposed in [122] overcomes the difficulties of convergence in MCMC methods. However all these algorithms work in offline mode and have high complexity. Online Bayesian CP detection algorithms were proposed in [120], [127]. Ideas from re-sampling algorithms for particle filters are applied to the forward/backward algorithm [122] to reduce the complexity without much loss in performance in [127]. On the other hand, [120] focused on casual predictive filtering. These algorithms assume the data within each segment is i.i.d., which is not true for EEG and ECoG data [128].

A.2 System Model and Notation

Let $x_{1:N} = (x_1,\ x_2, \cdots,\ x_N)^T$ denote the N data samples observed. Each data sample lies in $\mathbb{R}^d$, i.e., $x_n \in \mathbb{R}^d$, $\forall n = 1, 2, \cdots, N$. Note d is the number of electrodes used for recording EEG and ECoG data. Also, $x_{i:j}$ represents the data between time indices i and j, i.e., $(x_i,\ x_{i+1},\ \cdots,\ x_j)^T$. Let there be M change points (CPs) in this data sequence, denoted in increasing order by the time indices $\tau_1, \tau_2, \cdots, \tau_M$. By definition, $\tau_0 = 0$ and $\tau_{M+1} = N$. These change points imply, $x_{\tau_m+1:\tau_{m+1}}$ forms a segment of data drawn from some distribution, $\forall\, m = 0,\ 1,\ \cdots,\ M$ and that this underlying distribution is different in each segment. The objective is to find both the number of change points and their positions.

An auxiliary variable r_n, referred to as 'run-length' at time index n, is defined to help in inferring the change points. Run-length captures the time since the last change point [120]. Since $\tau_0 = 0$ is a change point, $r_1 = 0$. Also $r_n \in [0\ ,\ n-1]\,, \forall n$. Fig. A.1 plots some hypothetical time sample data x_n with $d = 1$ and the corresponding run-

length r_n. There are 2 change points τ_1 and τ_2 in this case and therefore r_n is 0 immediately after these 2 change points.

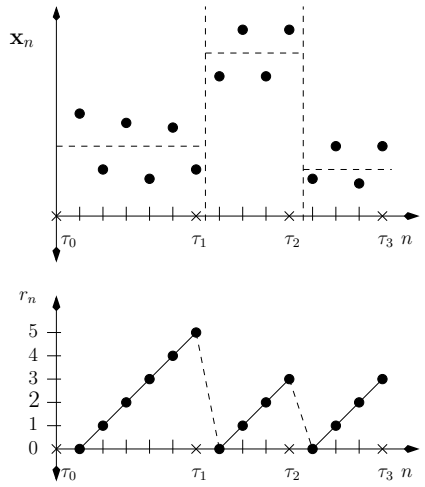


Figure A.1 : Example showing a sequence of data samples and the corresponding run-lengths

Given the run-length at a time instant, the run-length at the next time point can either go to 0 or increase by 1, depending on whether a change happens after this time instant or not. The relationship between r_n and $r_{n-1}, \forall\, n$ is given by

$$r_n = \begin{cases} 0 & , \text{ if } (n-1) \text{ is a change point} \\ r_{n-1} + 1 & , \text{ otherwise} \end{cases} \tag{A.1}$$

The conditional probability of r_n given the run-length at $(n-1)$ is denoted by $\mathrm{P}\left(r_n | r_{n-1}\right)$. In the following sub-sections, we define the prior probabilities on change points and the general model for likelihood of data within each segment, required for Bayesian inference in Sections A.3 and A.4.

A.2.1 Prior on Change Points

The change points are assumed to follow a Markov process, where the position of a change point is only dependent on the immediate preceding change point. The conditional probability of the k^{th} CP at some time index j given the $(k-1)^{th}$ change point at i is assumed to depend only on the distance between the change points [122, 126] for $k = 1,\ 2,\ \cdots, M$ and is given by

$$\mathrm{P}\left(\tau_k = j \mid \tau_{k-1} = i\right) = g\left(j - i\right),\ 0 \leq k-1 \leq i < j \leq N, \tag{A.2}$$

where $g\left(\cdot\right)$ is any discrete probability mass function over the set of natural numbers. Also note when $k = 1, i = 0$ in (A.2). The prior probability of a time index j being the k^{th} change point for $k = 1,\ 2,\ \cdots, M$, depends on the transition probability and is given by

$$\mathrm{P}\left(\tau_k = j\right) = \sum_{i=0}^{j-1} g\left(j - i\right) \cdot \mathrm{P}\left(\tau_{k-1} = i\right),\ j = 1,\ 2,\ \cdots,\ N. \tag{A.3}$$

Given $g\left(.\right)$, the transition probabilities for the run-length r_n given the run-length at the previous time instant is

$$\mathrm{P}\left(r_n = r | r_{n-1}\right) = \begin{cases} h\left(r_{n-1}+1\right) & ,\quad \text{if } r = 0 \\ 1 - h\left(r_{n-1}+1\right) & ,\quad \text{if } r = r_{n-1}+1 \end{cases} \tag{A.4}$$

where $h\left(r_{n-1}+1\right) = \frac{g(r_{n-1}+1)}{\sum_{i=r_{n-1}+1}^{\infty} g(i)}$, $\forall n = 2,\ 3,\ \cdots,\ N$. Also more generally, the probability of a segment whose length is atleast $(r+1)$ is given by

$$\mathrm{P}\left(r_n = r | r_{n-r} = 0\right) = \prod_{i=1}^{r} \left(1 - h\left(i\right)\right), \tag{A.5}$$

where $r \in [1, n-1]$ $\forall n = 2, 3, \cdots, N$. The CP prior information (A.2) is contained in the run-length transition probabilities (A.4) and the probability of a segment of some minimum length (A.5). The algorithms presented in Sections III and IV use (A.4) and (A.5), instead of (A.2).

A.2.2 Likelihood for a Data Segment

The data $x_{1:N}$ is assumed to satisfy the following property "given the positions of change points, the data in different segments is independent" [126]. These models are referred to as product partition models (PPM) [126]. Consider a data segment defined by the run-lengths $r_{n-r} = 0$ and $r_n = r$ at its two end-points. This data segment is assumed to be drawn from some distribution q, where q is an element in the set of some fixed number of distributions $\mathcal{Q}$. To get closed form expressions for likelihood, conjugate priors are defined on the parameters of q. Hyper-parameters are the parameters of the conjugate priors. The likelihood of a segment given a specific model $\mathrm{P}\left(x_{n-r:n} | r_{n-r} = 0, r_n = r, \mathrm{q}\right)$, is obtained by marginalizing over the parameters of q, but it still depends on the hyper-parameters. The explicit dependence of likelihood on hyper-parameters is dropped for notational convenience in the rest of the paper. The likelihood of the data within the segment coming from this set of distributions $\mathcal{Q}$, denoted by $\mathrm{P}\left(x_{n-r:n} | r_{n-r} = 0, r_n = r\right)$, is given by

$$\mathrm{P}\left(x_{n-r:n} | r_{n-r} = 0, r_n = r\right) = \sum_{\mathrm{q}} \mathrm{P}\left(x_{n-r:n} | r_{n-r} = 0, r_n = r, \mathrm{q}\right) \pi\left(\mathrm{q}\right), \tag{A.6}$$

where $\pi\left(\mathrm{q}\right)$ is the prior on the model space. Typically a uniform prior is used.

The closed form likelihood of a segment given a specific model (the first term inside the summation in (A.6)) is required to implement the algorithms described in

Sections III and IV. The algorithms described in Sections III and IV work with any model whose likelihood can be calculated or approximated. Linear regression models used to generate simulated data and also to model the EEG data are described in Section V.

A.3 Online Change Point Detection Algorithm

In this section, the online Bayesian CP detection algorithm is described. This algorithm extends the work done in [120] to the general case where the data within each segment is not i.i.d.

The positions and the number of change points are inferred from the posterior distribution of the run-length auxiliary variable. Specifically, the algorithm calculates $\mathrm{P}\left(r_n = r|x_{1:n}\right), \forall\, n = 1,\, 2,\, \cdots,\, N$ and $r = 0,\, 1,\, \cdots,\, n-1$ and infers change points from this posterior probability. This is done by calculating the joint probability of the run-length and the data observed upto that point and then by finding the desired conditional distribution. This algorithm has only one forward pass where each new data sample is used to compute a posterior probability of the run-length at that time instant. Change points can then be inferred from this posterior distribution. The proposed algorithm performs inference on a trellis shown in Fig. A.2, where each node represents the value of the auxiliary variable r_n. To illustrate this algorithm, we focus on computing the joint probability of r_4 with $x_{1:4}$ from the trellis in Fig. A.2. Going forward, node $r_4 = 0$ can only be reached via one of the 3 nodes, $r_3 = 0$, $r_3 = 1$ and $r_3 = 2$. The joint probability for $r_4 = 0$ is the weighted average of the product of the transition probabilities for moving from the 3 nodes at $n = 3$ to $r_4 = 0$ and the likelihood of the 4^{th} time instant being a segment on its own. The weights are the joint probabilities calculated at $n = 3$. Similarly, node $r_4 = 1$ can only be reached

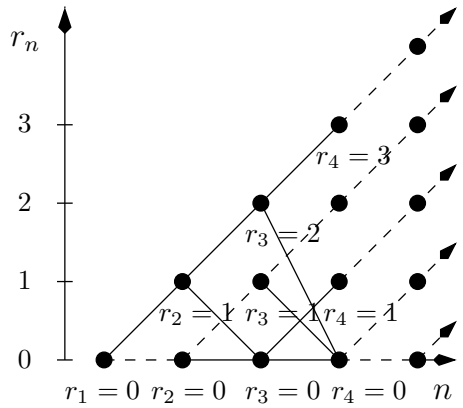


Figure A.2 : Online Bayesian CP detection algorithm on trellis

from nodes $r_2 = 0$ and $r_2 = 1$ via node $r_3 = 0$. The joint probability for $r_4 = 1$ is similarly calculated by the sum of the product of the joint probabilities at $n = 2$, the transition probability from r_2 to $r_3 = 0$ and from $r_3 = 0$ to $r_4 = 1$ and the likelihood of $x_{3:4}$ forming a segment. Finally, since there is only one path to reach node $r_4 = 3$, its joint probability is simply the product of transition probability from $r_1 = 0$ to $r_4 = 3$ and the likelihood of $x_{1:4}$ forming a segment. The posterior distribution of the auxiliary variable is then calculated from Baye's rule. In the following sub-section, the precise mathematical steps involved in calculating the posterior probability of run-length are derived.

A.3.1 Posterior Distribution of Run-length

First, we will compute the joint probability of the run-length and the data samples observed upto that time at each time instant. Using Markov property of change points (A.2) and the independence of data in different segments given the change points (PPMs), the joint probability of r_n and $x_{1:n}$ can be simplified. Since $\mathrm{P}(r_1 = 0) = 1$ at time index 1, $\mathrm{P}(r_1 = 0, x_{1:1}) = \mathrm{P}(x_{1:1}|r_1 = 0)$. For each time index $n = 2, 3, \cdots, N$

and $r = 0$, we have

$$\mathrm{P}\left(r_n = 0, x_{1:n}\right) = \sum_{i=0}^{n-2} \underbrace{\mathrm{P}\left(r_{n-1} = i, x_{1:n-1}\right)}_{\text{joint probability at } (n-1)} \cdot \underbrace{\mathrm{P}\left(r_n = 0 | r_{n-1} = i\right)}_{\text{transition probability}} \cdot \underbrace{\mathrm{P}\left(x_{n:n} | r_n = 0\right)}_{\text{data segment likelihood}}. \quad \text{(A.7)}$$

Similarly, for $n = 2,\ 3,\ \cdots,\ N$ and $r = n - 1$, we have

$$\mathrm{P}\left(r_n = n-1, x_{1:n}\right) = \mathrm{P}\left(x_{1:n} | r_1 = 0, r_n = n-1\right) \cdot \mathrm{P}\left(r_n = n-1 | r_1 = 0\right). \quad \text{(A.8)}$$

Finally, for $n = 3,\ \cdots,\ N$ and $r = 1,\ \cdots,\ n - 2$, we have

$$\begin{aligned}\mathrm{P}\left(r_n = r, x_{1:n}\right) = \sum_{i=0}^{n-r-2} & \mathrm{P}\left(r_{n-r-1} = i, x_{1:n-r-1}\right) \mathrm{P}\left(r_{n-r} = 0 | r_{n-r-1} = i\right) \\ & \cdot \mathrm{P}\left(x_{n-r:n} | r_{n-r} = 0, r_n = r\right) \mathrm{P}\left(r_n = r | r_{n-r} = 0\right). \quad \text{(A.9)}\end{aligned}$$

The posterior distribution of run-length is then calculated from the joint distribution using Bayes' rule.

$$\mathrm{P}\left(r_n = r | x_1 : n\right) = \frac{\mathrm{P}\left(r_n = r, x_{1:n}\right)}{\sum_{i=0}^{n-1} \mathrm{P}\left(r_n = i, x_{1:n}\right)}, \quad \text{(A.10)}$$

for $n = 2,\ \cdots,\ N$ and $r = 0,\ 1,\ \cdots,\ n - 1$.

A.3.2 Inferring the Change Points

The change points are inferred from the posterior probability calculated in (A.10) by back tracing from the final time index. This procedure is summarized below:

Set $m = 0$ and $\tau_0 = N$.

1. Find $r^\star = \arg\max_r \mathrm{P}\left(r_{\tau_m} = r | x_{1:\tau_m}\right)$.

2. Increment m by 1, i.e., $m \leftarrow m + 1$.

3. Add one more change point $\tau_m = \tau_{m-1} - r^\star - 1$.

4. If $\tau_m > 0$ go to step (1), else set M = m. The inferred change points are $(\tau_M, \tau_{M-1}, \cdots, \tau_1)$.

Calculating the likelihood of a segment is required to find the posterior of auxiliary variable by (A.7), (A.8), (A.9). The likelihood for the case of linear regression models is given in Section V.

A.4 Approximate Change Point Detection Algorithm

The number of different segment likelihoods computed by the CP detection algorithm described in the previous section grows as N^2, where N is the number of data samples. As a result the exact CP detection algorithm described in Section III has $\mathcal{O}(N^2)$ order complexity. Also the memory requirement to implement increases with N. In this section, we propose a simple approximation scheme that reduces the complexity from quadratic in N to linear in N.

The key idea is to compute the joint probability weights for only a fixed number of nodes N_p, instead of computing these weights at all $N(N-1)/2$ nodes in the trellis. The fixed number of nodes, N_p is constant with time. As a result the number of weights that are computed becomes linear with N. More specifically, at each time index n, we only retain $\mathrm{P}(r_n = r_i^\star | x_{1:n})$ where $r_i^\star \in [0, n-1]$ for $i = 1, 2, \cdots, N_p$. Note that $r_1^\star = 0$ and $r_{N_p}^\star = n-1$. For the next time instant $(n+1)$, the run-length can be in any one of $N_p + 1$ nodes with r values in the set $\mathcal{R} = \{0, r_1^\star + 1, r_2^\star + 1, \cdots, r_{N_p-1}^\star, n\}$. The joint probabilities at these $N_p + 1$ nodes are computed and the node with the smallest weight in the set $\mathcal{R} - \{0, n\}$ is discarded. Therefore at each time instant, the weights of only a fixed number of nodes are computed and stored

for further calculations. Change points are inferred from the posterior using the procedure described in Section A.3.2. Simulation results in Section A.6 demonstrate that this approximation works as well as the exact algorithm.

A.5 Likelihood Models

The two change point detection algorithms presented in Sections A.3 and A.4 are applicable for any data model. Both these algorithms use the likelihood of different segments of data in their computations (refer to (A.7), (A.8), (A.9)). In this section, the exact closed form likelihood expression for linear regression data models in white Gaussian noise is derived. The simulated data in Section VI and the real ECoG data in Section VII are modeled with this model.

The data in segment $x_{n-r:n}$ is modeled as

$$x_{n-r:n} = H\boldsymbol{\beta} + n, \tag{A.11}$$

where n is a $(r+1) \times 1$ i.i.d. Gaussian vector distributed with mean 0 and variance σ^2. H is a $(r+1) \times p$ matrix of basis functions and $\boldsymbol{\beta}$ the corresponding $p \times 1$ regression parameter vector. σ^2 is assumed to have an inverse Gamma prior with hyper-parameters $\nu/2$ and $\gamma/2$. $\boldsymbol{\beta}$ has a Gaussian prior with mean vector 0 and covariance matrix $\sigma^2 D$, where $D = \text{diag}\left(\delta_1^2, \delta_2^2, \cdots . \delta_p^2\right)$. The likelihood of a data

segment coming from this model conditioned on hyper-parameters is given by [129]

$$\mathrm{P}\left(x_{n-r:n}|r_{n-r}=0,r_n=r,q\right)=\pi^{(r+1)/2}\left(\frac{|M|}{|D|}\right)^{\frac{1}{2}}\frac{\gamma^{\nu/2}}{\left(\gamma+\|x_{n-r:n}\|_K^2\right)^{\frac{(r+1+\nu)}{2}}}\frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)},\text{ where} \tag{A.12}$$

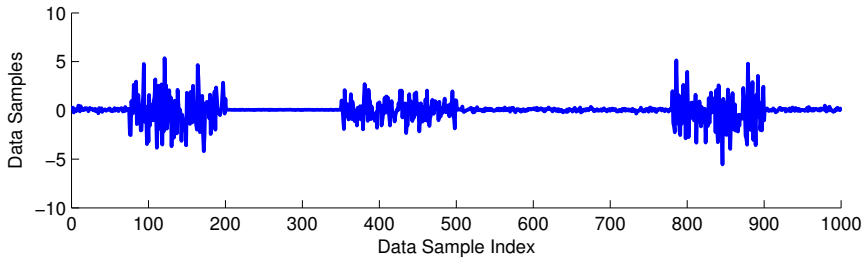
$$M=\left(H^TH+D^{-1}\right)^{-1},K=\left(I-HMH^T\right),\|x_{n-r:n}\|_K^2=x_{n-r:n}^TKx_{n-r:n}.$$

The likelihood of data within a segment coming from this model is obtained by substituting (A.12) into (A.6). The auxiliary variable posterior probabilities are calculated using the likelihood from (A.7), (A.8), (A.9) for the exact algorithm and from the procedure in Section A.4 for the approximate algorithm. Change points are then inferred using procedure described in Section A.3.2.

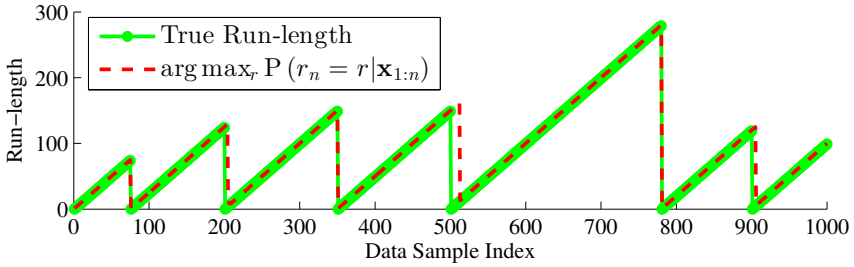
A.6 Performance on Simulated Data

In this section, the performance of the two change point detection algorithms in Sections A.3 and A.4 is evaluated on simulated data. The forward/backward CP detection algorithm in [122] is used as a benchmark to test the performance. The simulated data consists of 1000 data samples, with 6 change points shown in Fig. A.3a. The data in each segment is drawn from Gaussian distribution with mean 0 and some variances. The variance is different in each segment.

The data is assumed to be drawn from the model described in Section V. For this simulation, the values of hyper-parameters are $\nu=2,\gamma=2,\delta=1$. Also the change point transition probabilities (A.2) are modeled using a geometric distribution with parameter $\lambda=0.01$, since this leads to uniform probabilities on CP positions. Fig. A.3b plots the true value of run-length and the maximum of the posterior probability of the run-length for the exact Bayesian CP detection algorithm in Section III.



(a) Simulated 1000 data samples from Gaussian distribution with different variances



(b) True Run-Length and the Run-Length with Maximum Posterior given by Exact CP Bayesian detector in Section III

Figure A.3 : Performance Evaluation on Simulated Data

The slight offset between the two curves after change points is due to the time taken for the likelihood to drop because of the changing model. The back tracing described in Section III-B ensures that correct change points are detected. For this data, the exact algorithm, the approximate algorithm with $N_p = 10$ and the forward/backward algorithm [122] detect the same number of change points and the correct locations of all change points without any error.

A.7 Epileptic Activity Segmentation

The CP detection algorithms in Section III and IV are applied to ECoG data to identify segments of activity corresponding to the different states of epileptic brain. Data is ECoG recording from a patient with epilepsy. ECoG is recorded from 154

electrodes at 1000 Hz. Fig. A.4 shows a snapshot of 10 seconds of activity from 4 channels. In this work, we consider only shorter time windows of 500 samples from all 4 channels, where the channels correspond to 4 electrodes located between the temporal (T) and parietal (P) lobes of the brain. The channels are assumed to

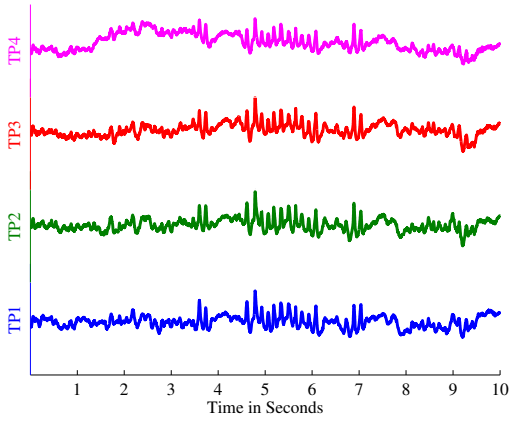


Figure A.4 : Snapshot of ECoG activity from 4 channels in a 10 second window

be independent and each channel is modeled using the model described in Section V. Also $\nu = 2, \gamma = 2, \delta = 1$ are taken as the values of the hyper-parameters for this simulation. Geometric distribution with $\lambda = 0.001$ is used to model the CP transition probabilities (A.2), since this leads to uniform probabilities on CP positions (A.3). The exact algorithm detected about 10 change points, where as the approximate one detected 9 change points. Further work needs to be done to extend the analysis to longer time windows and to incorporate the spatial and temporal correlations in EEG and ECoG data.

Appendix B

Appendix for Chapter 2

B.1 MI-in-frequency for Continuous-time Stochastic Processes

B.1.1 Proof of Equation (2.6)

We have from (2.4),

$$y(t) = \int_{-\infty}^{\infty} h_1(t-\tau)x(\tau)d\tau + \int_{-\infty}^{\infty} h_2(t-\tau)w(\tau)d\tau \tag{B.1}$$

We have from Theorem 3.1 and (B.1),

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{j2\pi\nu t} d\widetilde{Y}(\nu) &= \int_{-\infty}^{\infty} h_1(t-\tau) \int_{-\infty}^{\infty} e^{j2\pi\nu\tau} d\widetilde{X}(\nu)\, d\tau + \int_{-\infty}^{\infty} h_2(t-\tau) \int_{-\infty}^{\infty} e^{j2\pi\nu\tau} d\widetilde{W}(\nu)\, d\tau \\
&= \int_{-\infty}^{\infty} e^{j2\pi\nu t} \int_{-\infty}^{\infty} h_1(t-\tau) e^{-j2\pi\nu(t-\tau)} d\tau d\widetilde{X}(\nu) + \\
&\qquad \int_{-\infty}^{\infty} e^{j2\pi\nu t} \int_{-\infty}^{\infty} h_2(t-\tau) e^{-j2\pi\nu(t-\tau)} d\tau d\widetilde{W}(\nu) \qquad \text{(B.2)} \\
&= \int_{-\infty}^{\infty} e^{j2\pi\nu t} \left\{ H_1(j2\pi\nu)\, d\widetilde{X}(\nu) + H_2(j2\pi\nu)\, d\widetilde{W}(\nu) \right\}. \qquad \text{(B.3)} \\
&\implies d\widetilde{Y}(\nu) = H_1(j2\pi\nu)\, d\widetilde{X}(\nu) + H_2(j2\pi\nu)\, d\widetilde{W}(\nu).
\end{aligned}$$

B.1.2 Proof of Theorem 3.1

We will first prove that $\mathrm{MI}_{XY}(\nu_1, \nu_2)$ is zero, when $\nu_1 \neq \nu_2$, for the X and Y related by (2.4). Since the processes $X(t)$ and $W(t)$ are independent, their spectral processes are also independent. In addition, we also know from Theorem 3.2 that the spectral incre-

ments of the Gaussian process $X(t)$ are independent. It is clear from (2.6) that given $H_1(j2\pi\nu)$ and $H_2(j2\pi\nu)$, $\left[d\widetilde{Y}_R(\nu_2), d\widetilde{Y}_I(\nu_2)\right]$ is completely determined by the two-dimensional random vectors $\left[d\widetilde{X}_R(\nu_2), d\widetilde{X}_I(\nu_2)\right]$ and $\left[d\widetilde{W}_R(\nu_2), d\widetilde{W}_I(\nu_2)\right]$, both of which are independent of the two-dimensional random vector $\left[d\widetilde{X}_R(\nu_1), d\widetilde{X}_I(\nu_1)\right]$ when $\nu_1 \neq \nu_2$. This implies the mutual information between $\left[d\widetilde{Y}_R(\nu_2), d\widetilde{Y}_I(\nu_2)\right]$ and $\left[d\widetilde{X}_R(\nu_1), d\widetilde{X}_I(\nu_1)\right]$, which is defined as $\mathrm{MI}_{XY}(\nu_1, \nu_2)$, is zero.

We will now derive the analytical expression for $\mathrm{MI}_{XY}(\nu, \nu)$, for $\nu \neq 0$. Let $H_1(j2\pi\nu) = H_{1R}(j2\pi\nu) + jH_{1I}(j2\pi\nu)$ and $H_2(j2\pi\nu) = H_{2R}(j2\pi\nu) + jH_{2I}(j2\pi\nu)$. We can see from (2.5), (2.6) that

$$\left[d\widetilde{Y}_R(\nu), d\widetilde{Y}_I(\nu)\right] \sim \mathcal{N}\left(\mathbf{0}, \left(\tfrac{1}{2}f_X(\nu)\left|H_1(j2\pi\nu)\right|^2 + \tfrac{1}{2}f_W(\nu)\left|H_2(j2\pi\nu)\right|^2\right)\mathbf{I}\right), \tag{B.4}$$

where $\mathcal{N}$ represents Gaussian distribution, $\mathbf{0}$ is a two element zero vector and $\mathbf{I}$ is the 2×2 identity matrix. In addition,

$$\left[d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu), d\widetilde{Y}_R(\nu), d\widetilde{Y}_I(\nu)\right] \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right), \tag{B.5}$$

where $\sigma^2_{\widetilde{Y}}(\nu) = \left(f_X(\nu)\left|H_1(j2\pi\nu)\right|^2 + f_W(\nu)\left|H_2(j2\pi\nu)\right|^2\right)$, $\Sigma_{11} = \frac{1}{2}f_X(\nu)\mathbf{I}$, $\Sigma_{22} = \frac{1}{2}\sigma^2_{\widetilde{Y}}(\nu)\mathbf{I}$, $\mathbf{I}$ is the 2×2 identity matrix and $\mathbf{0}$ is a four element zero vector. In addition,

$$\Sigma_{12} = \Sigma_{21}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{2}f_X(\nu)H_{1R}(j2\pi\nu) & \frac{1}{2}f_X(\nu)H_{1I}(j2\pi\nu) \\ -\frac{1}{2}f_X(\nu)H_{1I}(j2\pi\nu) & \frac{1}{2}f_X(\nu)H_{1R}(j2\pi\nu) \end{bmatrix}.$$

Now, the MI between X and Y at frequency ν is given by

$$\begin{aligned}
\mathrm{MI}_{XY}(\nu,\nu) &= \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; \{d\widetilde{Y}_R(\nu), d\widetilde{Y}_I(\nu)\}\big) \\
&= \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_R(\nu)\big) + \\
&\qquad \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_I(\nu) \,|\, d\widetilde{Y}_R(\nu)\big) \qquad \text{(B.6)} \\
&= \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_R(\nu)\big) + \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_I(\nu)\big), \qquad \text{(B.7)}
\end{aligned}$$

where (B.6) follows from the chain rule of mutual information [56] and (B.7) follows because the real and imaginary parts of the spectral process of a Gaussian process are independent from Theorem 3.2. In addition, $\big[d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu), d\widetilde{Y}_R(\nu)\big]$ is a Gaussian distributed random vector with zero mean and covariance matrix Σ', which is easily obtained from (B.5). Since the mutual information between the components of a Gaussian random vector depends only on the determinants of the covariance matrices of the joint distribution and that of marginals [56], we can easily show that

$$\begin{aligned}
\mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_R(\nu)\big) &= \tfrac{1}{2}\log\frac{|\Sigma_{11}|\left(\frac{1}{2}\sigma^2_{\widetilde{Y}}\right)}{|\Sigma'|} \\
&= \tfrac{1}{2}\log\big(1+\tfrac{|H_1(j2\pi\nu)|^2 f_X(\nu)}{|H_2(j2\pi\nu)|^2 f_W(\nu)}\big). \qquad \text{(B.8)}
\end{aligned}$$

Similarly, we can also show that

$$\mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_I(\nu)\big) = \tfrac{1}{2}\log\big(1+\tfrac{|H_1(j2\pi\nu)|^2 f_X(\nu)}{|H_2(j2\pi\nu)|^2 f_W(\nu)}\big). \qquad \text{(B.9)}$$

From (B.7), (B.8) and (B.9), we have

$$\mathrm{MI}_{XY}(\nu,\nu) = 2 \times \mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_R(\nu)\big) = \log\big(1 + \tfrac{|H_1(j2\pi\nu)|^2 f_X(\nu)}{|H_2(j2\pi\nu)|^2 f_W(\nu)}\big). \tag{B.10}$$

At $\nu = 0$, MI-in-frequency between X and Y is equal to $\mathrm{I}\big(\{d\widetilde{X}_R(\nu), d\widetilde{X}_I(\nu)\}; d\widetilde{Y}_R(\nu)\big)$, since the imaginary part of Y is zero.

B.1.3 Relationship between MI-in-frequency and coherence

The coherence $R_{XY}(\nu) \in [0,1]$ between two processes X and Y related by (2.4) is given by

$$R_{XY}(\nu) = \tfrac{|f_{XY}(\nu)|^2}{f_X(\nu) f_Y(\nu)} = \tfrac{|H_1(j2\pi\nu)|^2 f_X(\nu)}{f_X(\nu)|H_1(j2\pi\nu)|^2 + f_W(\nu)|H_2(j2\pi\nu)|^2}.$$

$$\begin{aligned} \Rightarrow -\log(1 - R_{XY}(\nu)) &= \log\big(1 + \tfrac{|H_1(j2\pi\nu)|^2 f_X(\nu)}{|H_2(j2\pi\nu)|^2 f_W(\nu)}\big) \\ &= \mathrm{MI}_{XY}(\nu,\nu). \end{aligned} \tag{B.11}$$

B.2 MI-in-frequency for Discrete-time Stochastic Processes

B.2.1 Proof of Theorem 5.1

Now we consider two discrete-time Gaussian stochastic processes $X[n]$ and $Y[n]$ that are related by

$$y[n] = h_1[n] * x[n] + h_2[n] * w[n], \tag{B.12}$$

where $h_1[n]$ and $h_2[n]$ are the impulse responses of two discrete-time linear, time-invariant filters. (B.12) is the discrete-time equivalent of (2.4). It was shown in chapter 10 in [70] that mutual information between the discrete-time Gaussian stochastic processes $X[n]$ and $Y[n]$ is related to coherence according to

$$\mathrm{I}(X;Y) = -\int_0^{0.5} \log\left(1 - R_{XY}(\lambda)\right) d\lambda. \tag{B.13}$$

From (B.11) and (B.13), we have

$$\mathrm{I}(X;Y) = \int_0^{0.5} \mathrm{MI}_{XY}(\lambda,\lambda)\, d\lambda. \tag{B.14}$$

Appendix C

Appendix for Chapter 3

C.1 Proof of causal conditional entropy estimator

C.1.1 Proof of Lemma 3.1

First, we will prove the existence of $h(Y\|X)$. Since conditioning reduces differential entropy, we have

$$h(y[1]) \geq h(y[1]|X_1^1) \geq h(y[2]|Y_1^1, X_1^2) \geq \cdots \geq h\left(y[n]|Y_{n-J}^{n-1}, X_{n-K+1}^n\right) \geq \cdots\cdots \tag{C.1}$$

Therefore the sequence $h\left(y[n]|Y_{n-J}^{n-1}, X_{n-K+1}^n\right)$ is a non-increasing sequence that is upper bounded by $h(y[n])$. Also let $l = \max(J+1, K)$. Then for $n \geq l$,

$$h\left(y[n]|Y_1^{n-1}, X_1^n\right) = h\left(y[n]|Y_{n-J}^{n-1}, X_{n-K+1}^n\right) \tag{C.2}$$

$$= h\left(y[l]|Y_{l-J}^{l-1}, X_{l-K+1}^l\right), \tag{C.3}$$

where (C.2) is from the Markovian assumption and (C.3) is from the stationarity assumption. Note that (C.3) also implies the sequence $h\left(y[n]|Y_{n-J}^{n-1}, X_{n-K+1}^n\right)$ is lower bounded by $h\left(y[l]|Y_{l-J}^{l-1}, X_{l-K+1}^l\right)$. Let $a_n = h\left(y[n]|Y_1^{n-1}, X_1^n\right)$ and $b_N = \frac{1}{N}h\left(Y^N\|X^N\right) = \frac{1}{N}\sum_{n=1}^{N} a_n$. Since the $\lim_{N\to\infty} a_N$ exists, from Cesaro mean theorem [95] we have $h(Y\|X) = \lim_{N\to\infty} b_N$ also exists. The above proof can be easily modified to prove $h(Y)$ exists. Therefore $I(X \to Y) = h(Y) - h(Y\|X)$ also exists.

C.1.2 Proof of Lemma 3.2

$$\begin{aligned}
\tfrac{1}{N} h\left(Y^N \| X^N\right) &= \tfrac{1}{N} \sum_{n=1}^{N} h\left(y[n] | Y_{n-J}^{n-1}, X_{n-K+1}^{n}\right) && \text{(C.4)} \\
&= \tfrac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[-\log \mathrm{P}\left(y[l] | Y_{l-J}^{l-1}, X_{l-K+1}^{l}\right)\right] && \text{(C.5)} \\
&= \mathbb{E}\left[-\log \mathrm{P}\left(y[l] | Y_{l-J}^{l-1}, X_{l-K+1}^{l}\right)\right],
\end{aligned}$$

where (C.4) is from chain rule and Markovian assumption, and (C.5) is due to stationarity.

C.1.3 Proof of Theorem 3.1

Let $g_{J,K}\left(Y_{n-J}^{n}, X_{n-K+1)}^{n}\right) = -\log \mathrm{P}\left(y[n] | Y_{n-J}^{n-1}, X_{n-K+1}^{n}\right)$ be a fixed function over the states of the Markov chain $\left(Y_{n-J}^{n}, X_{n-K+1}^{n}\right)$. From the strong law of large numbers for Markov chains [100] which states that for a fixed function $g(.)$ over the states of the Markov chain, the sample mean will almost surely converge to the expected value as $N \to \infty$, we have,

$$\tfrac{1}{N} \sum_{n=1}^{N} g_{J,K}\left(Y_{n-J}^{n}, X_{n-K+1}^{n}\right) \xrightarrow{a.s.} \mathbb{E}\left[g_{J,K}\left(Y_{l-J}^{l}, X_{l-K+1}^{l}\right)\right]. \qquad \text{(C.6)}$$

We also have

$$\begin{aligned}
h\left(Y \| X\right) &= \lim_{N\to\infty} \tfrac{1}{N} h\left(Y^N \| X^N\right) && \text{(C.7)} \\
&= \lim_{N\to\infty} \mathbb{E}\left[g_{J,K}\left(Y_{l-J}^{l}, X_{l-K+1}^{l}\right)\right] && \text{(C.8)} \\
&= \mathbb{E}\left[g_{J,K}\left(Y_{l-J}^{l}, X_{l-K+1}^{l}\right)\right], && \text{(C.9)}
\end{aligned}$$

where (C.8) is from Lemma. 3.2. We have from (C.6), (C.9),

$$\hat{h}\left(Y\|X\right) = \frac{1}{N}\sum_{n=1}^{N} g_{J,K}\left(Y_{n-J}^{n}, X_{n-(K-1)}^{n}\right) \xrightarrow{a.s.} h\left(Y\|X\right). \tag{C.10}$$

C.2 Derivation of DI for Linear Two Node Network

Consider the MVAR model in section 3.5.1, described by (3.17). Here we will derive the DI in both directions between time-series X and Y for non-zero β_1, β_2. Appendix. C.2.3 considers the case when $(\beta_1, \beta_2) \in \{(1,0), (1,0)\}$.

C.2.1 DI from X to Y

For the system described by (3.17), the causal conditional entropy $h\left(Y\|X\right)$ is given by

$$h\left(Y\|X\right) = \lim_{N\to\infty} \frac{1}{N} h\left(Y^N\|X^N\right) = \frac{1}{2}\log\left(2\pi e\sigma_z^2\right), \tag{C.11}$$

because conditioned on $(x[n], x[n-1])$, the only uncertainty in $y[n]$ is due to the i.i.d Gaussian noise Z of variance σ_z^2 which is independent of X.

Now, from (3.17), we have $(y[1], y[2], \cdots, y[N])^T \sim \mathcal{N}\left(0, \Sigma_N\right)$, where $\Sigma_N = \delta M_N$ with $\delta = \beta_1\beta_2\sigma_x^2$. M_N is a tridiagonal matrix whose main diagonal elements are D and non-zero diagonal below and above the main diagonal are all 1. $D = \frac{\gamma}{\delta}$, where $\gamma = \left(\beta_1^2 + \beta_2^2\right)\sigma_x^2 + \sigma_z^2$. Upon further simplification using the tridiagonal matrix determinant from [107], we have

$$|\Sigma_N| = |\delta|^N \frac{\sinh((N+1)\lambda)}{\sinh\lambda}, \text{ where } \lambda = \cosh^{-1}\left(\frac{|D|}{2}\right). \tag{C.12}$$

The unconditioned entropy of Y is now given by

$$h(Y) = \lim_{N\to\infty} \tfrac{1}{2N} \log\left((2\pi e)^N |\Sigma_N|\right) = \tfrac{1}{2}\log(2\pi e|\delta|) + \tfrac{1}{2}\lambda, \tag{C.13}$$

obtained by expanding the hyperbolic sinh function in the determinant $|\Sigma_N|$ in terms of exponentials and some basic algebraic manipulations. Now, from (C.11) and (C.13), we have

$$I(X \to Y) = \tfrac{1}{2}\log\left(\tfrac{|\beta_1\beta_2|\sigma_x^2}{\sigma_z^2}\right) + \tfrac{1}{2}\cosh^{-1}\left(\tfrac{\left(\beta_1^2+\beta_2^2\right)\sigma_x^2+\sigma_z^2}{2|\beta_1\beta_2|\sigma_x^2}\right).$$

C.2.2 DI from Y to X

The causal conditional entropy, $h(X\|Y)$ is given by

$$\begin{aligned} h(X\|Y) &= \lim_{N\to\infty} \tfrac{1}{N}\sum_{n=1}^{N} h(x[n]|x[n-1], y[n]) && \text{(C.14)}\\ &= \lim_{N\to\infty} \tfrac{1}{N}\sum_{n=1}^{N} \{h(x[n], y[n], x[n-1]) - h(x[n-1], y[n])\} \\ &= \lim_{N\to\infty} \tfrac{1}{N}\sum_{n=1}^{N} \left\{\tfrac{1}{2}\log(2\pi e|\Phi_1|) - \tfrac{1}{2}\log(2\pi e|\Phi_2|)\right\} \\ &= \tfrac{1}{2}\log\left(2\pi e \tfrac{\sigma_x^2\sigma_z^2}{\beta_1^2\sigma_x^2+\sigma_z^2}\right), && \text{(C.15)} \end{aligned}$$

where Φ_1 and Φ_2 are the appropriate covariance matrices. The reason for (C.14) is that conditioned on $x[n-1]$ and $y[n]$, $x[n]$ is independent of the other past samples of X and Y. Since X is drawn from i.i.d. Gaussian distribution with mean zero and variance σ_x^2, the unconditional entropy of X is given by $h(X) = \tfrac{1}{2}\log(2\pi e\sigma_x^2)$.

Therefore, the DI from Y to X is

$$I\left(Y \to X\right) = h\left(X\right) - h\left(X \| Y\right) = \tfrac{1}{2}\log\left(1 + \tfrac{\beta_1^2\sigma_x^2}{\sigma_z^2}\right). \tag{C.16}$$

C.2.3 Special cases

Consider the system in (3.17) with $\beta_1 = 1, \beta_2 = 0$. For this system, $y[n]$ are i.i.d. Gaussian distributed with mean zero and variance $(\sigma_x^2 + \sigma_z^2)$. Therefore the differential entropy of Y_1^N is given by $h\left(Y^N\right) = \frac{N}{2}\log\left(2\pi e\left(\sigma_x^2 + \sigma_z^2\right)\right)$. Also the joint differential entropy of $x[n]$ and $y[n]$ is

$$h\left(x[n], y[n]\right) = \tfrac{1}{2}\log\left(2\pi e \begin{vmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_z^2 \end{vmatrix}\right) = \tfrac{1}{2}\log\left(2\pi e \sigma_x^2 \sigma_z^2\right). \tag{C.17}$$

$$\implies h\left(Y^N \| X^N\right) = \sum_{n=1}^{N} h\left(y[n]|x[n]\right) = \sum_{n=1}^{N}\left(h\left(x[n], y[n]\right) - h\left(x[n]\right)\right) = \tfrac{N}{2}\log\left(2\pi e\sigma_z^2\right).$$

Therefore the directed information from X to Y is given by

$$I\left(X \to Y\right) = \lim_{N\to\infty}\left(h\left(Y^N\right) - h\left(Y^N \| X^N\right)\right) = \tfrac{1}{2}\log\left(1 + \tfrac{\sigma_x^2}{\sigma_z^2}\right). \tag{C.18}$$

The DI from Y to X can be similarly derived.

Now, consider the system in (3.17) with $\beta_1 = 0, \beta_2 = 1$. For this system, the DI from X to Y is computed by following the approach described above. Let us derive $I\left(Y \to X\right)$. The causal conditional entropy of X^N given Y^N is given by

$$h\left(X^N \| Y^N\right) = \sum_{n=1}^{N} h\left(x[n]|X^{n-1}, Y^N\right) = \sum_{n=1}^{N} h\left(x[n]\right) = h\left(X^N\right), \tag{C.19}$$

since $x[n]$ does not depend on the past samples of Y. Therefore, the DI from Y to X

is zero, i.e., $I(Y \to X) = 0$.

Bibliography

[1] S. M. Kay, "Statistical signal processing," *Estimation Theory*, vol. 1, 1993.

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[3] S. L. Lauritzen, *Graphical models*, vol. 17. Clarendon Press, 1996.

[4] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[5] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[6] S. Zhou, J. Lafferty, and L. Wasserman, "Time varying undirected graphs," *Machine Learning*, vol. 80, no. 2-3, pp. 295–319, 2010.

[7] M. Kolar, L. Song, A. Ahmed, and E. P. Xing, "Estimating time-varying networks," *The Annals of Applied Statistics*, pp. 94–123, 2010.

[8] H. V. Poor and O. Hadjiliadis, *Quickest detection*, vol. 40. Cambridge University Press Cambridge, 2009.

[9] R. Malladi, G. P. Kalamangalam, and B. Aazhang, "Online bayesian change point detection algorithms for segmentation of epileptic activity," in *Asilomar Conf. on Signals, Systems and Computers*, 2013.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[11] K. J. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[12] K. P. Murphy, *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[13] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 1961.

[14] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 1. Springer, 2006.

[15] R. Malladi, D. H. Johnson, G. Kalamangalam, N. Tandon, and B. Aazhang, "Measuring cross-frequency coupling using mutual information and its application to epilepsy," in *Cosyne Abstracts*, (Salt Lake City, USA), 2017.

[16] R. Malladi, D. Johnson, and B. Aazhang, "Data-Driven estimation of mutual information between dependent data," *submitted to IEEE International Symposium on Information Theory (ISIT)*, June 2017.

[17] R. Malladi, D. H. Johnson, G. Kalamangalam, N. Tandon, and B. Aazhang, "Mutual information in frequency and its application to epilepsy," *submitted to IEEE Transactions on Signal Processing*, 2017.

[18] R. Malladi, G. Kalamangalam, N. Tandon, and B. Aazhang, "Identifying the epileptogenic zone using directed information," in *Abstract presented at Soc. for Neuroscience (SfN)*, 2014.

[19] R. Malladi, G. P. Kalamangalam, N. Tandon, and B. Aazhang, "Inferring causal connectivity in epileptogenic zone using directed information," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 822–826, 2015.

[20] R. Malladi, G. Kalamangalam, N. Tandon, and B. Aazhang, "Identifying seizure onset zone from the causal connectivity inferred using directed information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 1267–1283, Oct 2016.

[21] F. Rosenow and H. Lüders, "Presurgical evaluation of epilepsy," *Brain*, vol. 124, no. 9, pp. 1683–1700, 2001.

[22] G. K. Bergey, M. J. Morrell, E. M. Mizrahi, A. Goldman, D. King-Stephens, D. Nair, S. Srinivasan, B. Jobst, R. E. Gross, D. C. Shields, *et al.*, "Long-term treatment with responsive brain stimulation in adults with refractory partial seizures," *Neurology*, vol. 84, pp. 810–817, 2015.

[23] U. Gleissner, R. Sassen, M. Lendt, H. Clusmann, C. Elger, and C. Helmstaedter, "Pre-and postoperative verbal memory in pediatric patients with temporal lobe epilepsy," *Epilepsy research*, vol. 51, no. 3, pp. 287–296, 2002.

[24] C. H. Halpern, U. Samadani, B. Litt, J. L. Jaggi, and G. H. Baltuch, "Deep brain stimulation for epilepsy," *Neurotherapeutics*, vol. 5, no. 1, 2008.

[25] M. Rodriguez-Oroz, J. Obeso, A. Lang, J.-L. Houeto, P. Pollak, S. Rehncrona, J. Kulisevsky, A. Albanese, J. Volkmann, M. Hariz, *et al.*, "Bilateral deep brain stimulation in parkinson's disease: a multicentre study with 4 years follow-up," *Brain*, vol. 128, no. 10, pp. 2240–2249, 2005.

[26] S. Sunderam, B. Gluckman, D. Reato, and M. Bikson, "Toward rational design of electrical stimulation strategies for epilepsy control," *Epilepsy & Behavior*, vol. 17, no. 1, pp. 6–22, 2010.

[27] E. Krook-Magnuson and I. Soltesz, "Beyond the hammer and the scalpel: selective circuit control for the epilepsies," *Nature Neuroscience*, vol. 18, no. 3, pp. 331–338, 2015.

[28] H. O. Luders, I. Najm, D. Nair, P. Widdess-Walsh, and W. Bingman, "The epileptogenic zone: general principles," *Epileptic Disorders*, vol. 8, 2006.

[29] M. D. Holmes and D. M. Tucker, "Identifying the epileptic network," *Frontiers in Neurology*, vol. 4, p. 84, 2013.

[30] H. Stefan and F. H. Lopes Da Silva, "Epileptic neuronal networks: methods of identification and clinical relevance.," *Frontiers in neurology*, vol. 4, p. 8, 2013.

[31] M. Leite, A. Leal, and P. Figueiredo, "Transfer function between eeg and bold signals of epileptic activity," *Frontiers in Neurology*, vol. 4, p. 1, 2013.

[32] K. Weaver, W. Chaovalitwongse, E. Novotny, A. Poliakov, T. Grabowski, and J. Ojemann, "Local functional connectivity as a pre-surgical tool for seizure focus identification in non-lesion, focal epilepsy," *Frontiers in Neurology*, vol. 4, 2013.

[33] W. T. Kerr, S. T. Nguyen, A. Y. Cho, E. P. Lau, D. H. Silverman, P. K. Douglas, N. M. Reddy, A. Anderson, J. Bramen, N. Salamon, *et al.*, "Computer-aided diagnosis and localization of lateralized temporal lobe epilepsy using interictal fdg-pet," *Frontiers in Neurology*, vol. 4, 2013.

[34] M. Yamazaki, D. M. Tucker, M. Terrill, A. Fujimoto, and T. Yamamoto, "Dense array eeg source estimation in neocortical epilepsy," *Frontiers in Neurology*, vol. 4, 2013.

[35] J. Zhang, W. Cheng, Z. Wang, Z. Zhang, W. Lu, G. Lu, and J. Feng, "Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy," *PloS one*, vol. 7, no. 5, p. e36733, 2012.

[36] H. Blumenfeld, K. A. McNally, S. D. Vanderhill, A. L. Paige, R. Chung, K. Davis, A. D. Norden, R. Stokking, C. Studholme, E. J. Novotny, *et al.*, "Positive and negative network correlations in temporal lobe epilepsy," *Cerebral cortex*, vol. 14, no. 8, pp. 892–902, 2004.

[37] E. H. Bertram, "Neuronal circuits in epilepsy: do they matter?," *Experimental neurology*, vol. 244, pp. 67–74, 2013.

[38] J. Gotman, C. Grova, A. Bagshaw, E. Kobayashi, Y. Aghakhani, and F. Dubeau, "Generalized epileptic discharges show thalamocortical activation and suspension of the default state of the brain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 42, pp. 15236–15240, 2005.

[39] B. C. Bernhardt, N. Bernasconi, H. Kim, and A. Bernasconi, "Mapping thalamocortical network pathology in temporal lobe epilepsy," *Neurology*, vol. 78, no. 2, pp. 129–136, 2012.

[40] J. G. Milton, "Epilepsy as a dynamic disease: a tutorial of the past with an eye to the future," *Epilepsy & behavior*, vol. 18, no. 1, pp. 33–44, 2010.

[41] K. J. Friston, "Functional and effective connectivity in neuroimaging: a synthesis," *Human Brain Mapping*, vol. 2, no. 1-2, 1994.

[42] B. Horwitz, "The elusive concept of brain connectivity," *Neuroimage*, vol. 19, no. 2, pp. 466–470, 2003.

[43] K. J. Friston, "Functional and effective connectivity: a review," *Brain connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[44] E. W. Lang, A. Tomé, I. R. Keck, J. Górriz-Sáez, and C. Puntonet, "Brain connectivity analysis: a short survey," *Computational Intelligence and Neuroscience*, vol. 2012, p. 8, 2012.

[45] P. J. Franaszczuk and G. K. Bergey, "Application of the directed transfer function method to mesial and lateral onset temporal lobe seizures," *Brain Topography*, vol. 11, no. 1, pp. 13–21, 1998.

[46] S. Sabesan, L. B. Good, K. S. Tsakalis, A. Spanias, D. M. Treiman, and L. D. Iasemidis, "Information flow and application to epileptogenic focus localization from intracranial EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 3, pp. 244–253, 2009.

[47] F. Wendling, P. Chauvel, A. Biraben, and F. Bartolomei, "From intracerebral EEG signals to brain connectivity: identification of epileptogenic networks in partial epilepsy," *Frontiers in Systems Neuroscience*, 2010.

[48] C. Wilke, W. Van Drongelen, M. Kohrman, and B. He, "Neocortical seizure foci localization by means of a directed transfer function method," *Epilepsia*, vol. 51, no. 4, pp. 564–572, 2010.

[49] F. Panzica, G. Varotto, F. Rotondi, R. Spreafico, and S. Franceschetti, "Identification of the epileptogenic zone from stereo-EEG signals: a connectivity-graph theory approach," *Frontiers in Neurology*, 2013.

[50] P. van Mierlo, M. Papadopoulou, E. Carrette, P. Boon, S. Vandenberghe, K. Vonck, and D. Marinazzo, "Functional brain connectivity from EEG in epilepsy: Seizure prediction and epileptogenic focus localization," *Progress in neurobiology*, vol. 121, pp. 19–35, 2014.

[51] F. Pittau and S. Vulliemoz, "Functional brain networks in epilepsy: recent advances in noninvasive mapping," *Current opinion in neurology*, vol. 28, no. 4, pp. 338–343, 2015.

[52] K. J. Blinowska, "Review of the methods of determination of directed connectivity from multichannel data," *Medical & Biological Eng. & Computing*, vol. 49, no. 5, 2011.

[53] S. P. Burns, S. Santaniello, R. B. Yaffe, C. C. Jouny, N. E. Crone, G. K. Bergey, W. S. Anderson, and S. V. Sarma, "Network dynamics of the brain and influence of the epileptic seizure onset zone," *Proceedings of the National Academy of Sciences*, vol. 111, no. 49, 2014.

[54] "Focus on epilepsy," *Nature Neuroscience*, vol. 18, p. 317, Feb. 2015.

[55] C. E. Shannon, "A mathematical theory of communication," 1948.

[56] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[57] D. R. Brillinger, "Second-order moments and mutual information in the analysis of time series," *Recent Advances in Statistical Methods*, pp. 64–76, 2002.

[58] Q. Wang, S. R. Kulkarni, and S. Verdú, "Universal estimation of information measures for analog sources," *Foundations and Trends in Communications and Information Theory*, vol. 5, pp. 265–353, 2009.

[59] R. T. Canolty, E. Edwards, S. S. Dalal, M. Soltani, S. S. Nagarajan, H. E. Kirsch, M. S. Berger, N. M. Barbaro, and R. T. Knight, "High gamma power is phase-locked to theta oscillations in human neocortex," *science*, vol. 313, no. 5793, pp. 1626–1628, 2006.

[60] R. T. Canolty and R. T. Knight, "The functional role of cross-frequency coupling," *Trends in cognitive sciences*, vol. 14, no. 11, pp. 506–515, 2010.

[61] L. Faes and G. Nollo, *Multivariate frequency domain analysis of causal interactions in physiological time series*. INTECH Open Access Publisher, 2011.

[62] J. Aru, J. Aru, V. Priesemann, M. Wibral, L. Lana, G. Pipa, W. Singer, and R. Vicente, "Untangling cross-frequency coupling in neuroscience," *Current opinion in neurobiology*, vol. 31, pp. 51–61, 2015.

[63] R. Pascual-Marqui, P. Faber, T. Kinoshita, Y. Kitaura, K. Kochi, P. Milz, K. Nishida, and M. Yoshimura, "The dual frequency rv-coupling coefficient: a novel measure for quantifying cross-frequency information transactions in the brain," *arXiv preprint arXiv:1603.05343*, 2016.

[64] H. Cramér and M. Leadbetter, *Stationary and related stochastic processes: sample function properties and their applications*. Wiley series in probability and mathematical statistics. Tracts on probability and statistics, Wiley, 1967.

[65] H. J. Larson and B. O. Shubert, *Probabilistic models in engineering sciences*, vol. 2. Wiley, 1979.

[66] D. R. Brillinger, *Time series: data analysis and theory*, vol. 36. Siam, 2001.

[67] S. Khan, S. Bandyopadhyay, A. R. Ganguly, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, "Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data," *Physical Review E*, vol. 76, no. 2, p. 026209, 2007.

[68] E. Schaffernicht, R. Kaltenhaeuser, S. S. Verma, and H.-M. Gross, "On estimating mutual information for feature selection," in *International Conference on Artificial Neural Networks*, pp. 362–367, Springer, 2010.

[69] A. C. Onslow, R. Bogacz, and M. W. Jones, "Quantifying phase–amplitude coupling in neuronal network oscillations," *Progress in biophysics and molecular biology*, vol. 105, no. 1, pp. 49–57, 2011.

[70] M. S. Pinsker, "Information and information stability of random variables and processes," 1960.

[71] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *et al.*, *Discrete-time signal processing*, vol. 2. Prentice hall Englewood Cliffs, NJ, 1989.

[72] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Progress in Neurobiology*, vol. 77, no. 1, pp. 1–37, 2005.

[73] R. Salvador, A. Martinez, E. Pomarol-Clotet, J. Gomar, F. Vila, S. Sarro, A. Capdevila, and E. Bullmore, "A simple view of the brain through a frequency-specific functional connectivity measure," *Neuroimage*, vol. 39, no. 1, pp. 279–289, 2008.

[74] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[75] T. Duong *et al.*, "ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R," *J. of Statistical Software*, vol. 21, no. 7, pp. 1–16, 2007.

[76] D. Dvorak and A. A. Fenton, "Toward a proper estimation of phase–amplitude coupling in neural oscillations," *Journal of neuroscience methods*, vol. 225, pp. 42–56, 2014.

[77] J. I. Berman, J. McDaniel, S. Liu, L. Cornew, W. Gaetz, T. P. Roberts, and J. C. Edgar, "Variable bandwidth filtering for improved sensitivity of cross-frequency coupling metrics," *Brain connectivity*, vol. 2, no. 3, pp. 155–163, 2012.

[78] S. L. Bressler and A. K. Seth, "Wiener–Granger causality: a well established methodology," *Neuroimage*, vol. 58, pp. 323–329, 2011.

[79] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. of the Econometric Soc.*, 1969.

[80] T. Schreiber, "Measuring information transfer," *Physical Review Lett.*, vol. 85, no. 2, 2000.

[81] K. Lehnertz, "Epilepsy and nonlinear dynamics," *J. of biological physics*, vol. 34, no. 3-4, pp. 253–266, 2008.

[82] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J. of Computational Neuroscience*, vol. 30, no. 1, 2011.

[83] K. So, A. C. Koralek, K. Ganguly, M. C. Gastpar, and J. M. Carmena, "Assessing functional connectivity of neural ensembles using directed information," *J. of Neural Eng.*, vol. 9, no. 2, 2012.

[84] N. Soltani and A. Goldsmith, "Directed information between connected leaky integrate-and-fire neurons," in *IEEE Int. Symp. on Inform. Theory (ISIT)*, pp. 1291–1295, 2014.

[85] H. Marko, "The bidirectional communication theory–a generalization of information theory," *IEEE Trans. Commun.*, vol. 21, no. 12, 1973.

[86] J. Massey, "Causality, feedback and directed information," in *Int. Symp. on Inform. Theory Applications (ISITA)*, 1990.

[87] G. Kramer, *Directed information for channels with feedback*. PhD thesis, ETH Zürich, 1998.

[88] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[89] Y. Liu, *Directed Information for Complex Network Analysis from Multivariate Time Series*. PhD thesis, Michigan State University, East Lansing, MI, USA, 2012.

[90] Y. Liu and S. Aviyente, "The relationship between transfer entropy and directed information," in *IEEE Statistical Signal Process. Workshop (SSP)*, pp. 73–76, Aug 2012.

[91] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Inference of biologically relevant gene influence networks using the directed information criterion," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 2, 2006.

[92] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, 2009.

[93] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, 2011.

[94] P.-O. Amblard and O. J. Michel, "On directed information theory and Granger causality graphs," *J. of Computational Neuroscience*, vol. 30, no. 1, 2011.

[95] T. Cover and J. Thomas, *Elements of information theory*. Wiley, 2006.

[96] L. Barnett and A. K. Seth, "The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference," *J. of Neuroscience Methods*, vol. 223, pp. 50–68, 2014.

[97] C. Diks and J. DeGoede, "A general nonparametric bootstrap test for Granger causality," in *Global Analysis of Dynamical Systems*, 2001.

[98] D. N. Politis and J. P. Romano, "The stationary bootstrap," *J. of the Amer. Statistical association*, vol. 89, no. 428, pp. 1303–1313, 1994.

[99] K. J. Blinowska and M. Kamiński, "Multivariate signal analysis by parametric models," *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, p. 373, 2006.

[100] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Cambridge University Press, 2009.

[101] N.-Z. Shi and J. Tao, *Statistical hypothesis testing: theory and methods*. World Scientific, 2008.

[102] S. M. Kay, *Fundamentals of statistical signal processing : estimation theory.* Prentice-Hall, 2010.

[103] A. J. Izenman, "Review papers: recent developments in nonparametric density estimation," *J. of the Amer. Statistical Assoc.*, vol. 86, no. 413, pp. 205–224, 1991.

[104] P. D. Grünwald, *The minimum description length principle.* MIT press, 2007.

[105] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, 1991.

[106] D. Wied and R. Weißbach, "Consistency of the kernel density estimator: a survey," *Statistical Papers*, vol. 53, no. 1, pp. 1–21, 2012.

[107] G. Hu and R. O'Connell, "Analytical inversion of symmetric tridiagonal matrices," *J. of Physics A: Math. and General*, vol. 29, no. 7, 1996.

[108] K. Ishiguro, N. Otsu, M. Lungarella, and Y. Kuniyoshi, "Comparison of nonlinear Granger causality extensions for low-dimensional systems," *Physical Review E*, vol. 77, no. 3, p. 036217, 2008.

[109] C. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *arXiv preprint arXiv:1204.2003*, 2012.

[110] C. Wilke, G. Worrell, and B. He, "Graph analysis of epileptogenic networks in human partial epilepsy," *Epilepsia*, vol. 52, no. 1, 2011.

[111] K. Edakawa, T. Yanagisawa, H. Kishima, R. Fukuma, S. Oshino, H. M. Khoo, M. Kobayashi, M. Tanaka, and T. Yoshimine, "Detection of epileptic seizures using phase–amplitude coupling in intracranial electroencephalography," *Scientific Reports*, vol. 6, p. 25422, 2016.

[112] S. A. Weiss, A. Lemesiou, R. Connors, G. P. Banks, G. M. McKhann, R. R. Goodman, B. Zhao, C. G. Filippi, M. Nowell, R. Rodionov, *et al.*, "Seizure localization using ictal phase-locked high gamma a retrospective surgical outcome study," *Neurology*, vol. 84, no. 23, pp. 2320–2328, 2015.

[113] M. Guirgis, Y. Chinvarun, M. del Campo, P. L. Carlen, and B. L. Bardakjian, "Defining regions of interest using cross-frequency coupling in extratemporal lobe epilepsy patients," *Journal of neural engineering*, vol. 12, no. 2, p. 026011, 2015.

[114] C. Tonini, E. Beghi, A. T. Berg, G. Bogliun, L. Giordano, R. W. Newton, A. Tetto, E. Vitelli, D. Vitezic, and S. Wiebe, "Predictors of epilepsy surgery outcome: a meta-analysis," *Epilepsy research*, vol. 62, no. 1, pp. 75–87, 2004.

[115] P. Mierlo, E. Carrette, H. Hallez, R. Raedt, A. Meurs, S. Vandenberghe, D. Roost, P. Boon, S. Staelens, and K. Vonck, "Ictal-onset localization through connectivity analysis of intracranial EEG signals in patients with refractory epilepsy," *Epilepsia*, vol. 54, no. 8, pp. 1409–1418, 2013.

[116] C. P. Warren, S. Hu, M. Stead, B. H. Brinkmann, M. R. Bower, and G. A. Worrell, "Synchrony in normal and focal epileptic brain: the seizure onset zone is functionally disconnected," *J. of neurophysiology*, vol. 104, no. 6, pp. 3530–3539, 2010.

[117] C. J. Stam, G. Nolte, and A. Daffertshofer, "Phase lag index: assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources," *Human brain mapping*, vol. 28, no. 11, pp. 1178–1193, 2007.

[118] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[119] R. S. Fisher, "Neurostimulation for epilepsy: do we know the best stimulation parameters?," *Epilepsy Currents*, vol. 11, no. 6, 2011.

[120] R. P. Adams and D. J. C. MacKay, "Bayesian online changepoint detection," tech. rep., University of Cambridge, Cambridge, UK, 2007.

[121] N. Seshadri and C. E. W. Sundberg, "List viterbi decoding algorithms with applications," *IEEE Transactions on Communications*, vol. 42, no. 234, pp. 313–323, 1994.

[122] P. Fearnhead, "Exact bayesian curve fitting and signal segmentation," *IEEE Transactions on Signal Processing*, vol. 53, no. 6, pp. 2160–2166, 2005.

[123] S. B. Wilson and R. Emerson, "Spike detection: a review and comparison of algorithms," *Clinical Neurophysiology*, vol. 113, no. 12, pp. 1873–1881, 2002.

[124] P. J. Green, "Reversible jump Markov chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, vol. 82, pp. 711–732, Dec. 1995.

[125] E. Punskaya, C. Andrieu, A. Doucet, and W. J. Fitzgerald, "Bayesian curve fitting using MCMC with applications to signal segmentation," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 747–758, 2002.

[126] D. Barry and J. A. Hartigan, "A bayesian analysis for change point problems," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 309–319, 1993.

[127] P. Fearnhead and Z. Liu, "On-line inference for multiple changepoint problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, 2007.

[128] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 14, pp. 77 – 94, 1992.

[129] X. Xuan, "Bayesian inference on change point problems," Master's thesis, University of British Columbia, 2007.